الإحصاء
مع برنامج ستاتا

تأليف
لورنس سي هاميلتون
Lawrence C.Hamilton

ترجمة
د. رمضان مفتاح الفيتوري

مراجعة
أ. الفيتوري مفتاح الفيتوري



الإحصاء

مع برنامج ستاتا

الإحصاء

مع برنامج ستاتا

تأليف

لورنس سي هاميلتون
Lawrence C. Hamilton

ترجمة

د. رمضان مفتاح الفيتوري
كلية الاقتصاد - جامعة عمر المختار

مراجعة

أ. الفيتوري مفتاح الفيتوري
كلية الهندسة - جامعة عمر المختار



المملكة العربية السعودية - الرياض - هاتف: 4658523 – 4647531 + (0096611)

ص. ب 10720 – الرمز البريدي: 11443 - فاكس: 4657939 + (0096611)

الطبعة الإنجليزية:

STATISTICS WITH STATA

By. Lawrence C. Hamilton

ردمك : 9960 - 24 - 735 - X

© دار المريخ للنشر

المملكة العربية السعودية، الرياض ، 1436هـ/2015م

المملكة العربية السعودية - الرياض - ص . ب: 10720 الرمز البريدي: 11443

هاتف: 4658523 / 4647531 فاكس: 4657939 + (0096611)

البريد الالكتروني : *Email : mars@marspub1.com*

التوزيع داخل جمهورية مصر العربية والسودان وشمال أفريقيا :

دار المريخ للنشر بالقاهرة – 4 شارع الفرات – المهندسين – الجيزة – الرمز البريدي: 12411

هاتف: 37609971/ 33376579 فاكس: 37609457 + (00202)

البريد الالكتروني: *Email : marspub2002@yahoo.com*



محتويات الكتاب

صفحة

الفصل الأول

ستاتا ومصادر البيانات

الفصل الثاني

إدارة البيانات

صفحة

صفحة

الفصل الرابع

بيانات الدراسات الاستقصائية

الفصل الخامس

الملخصات الإحصائية والجداول

صفحة

الفصل السادس

تحليل التباين وطرق المقارنة الأخرى

الفصل السابع

تحليل الانحدار الخطي

صفحة

الفصل الثامن

طرق الانحدار المتقدمة

الفصل التاسع

الانحدار اللوغاريتمي

الفصل العاشر

نماذج عد الأحداث والبقاء

صفحة

الفصل الحادي عشر

تحليل المكونات الرئيسية

التحليل العاملي والتحليل العنقودي

الفصل الثاني عشر

تحليل السلاسل الزمنية

صفحة

الفصل الثالث عشر

صياغة نماذج التأثيرات المختلطة والمستويات المتعددة

الفصل الرابع عشر

مقدمة في البرمجة

صفحة

مقدمة المؤلف

"الإحصاء مع ستاتا".. مخصص للطلبة والباحثين العمليين، ويهدف إلى سد الفجوة بين كُتب الإحصاء وأدلة استخدام ستاتا. وليس دور هذا الكتاب إعطاء توضيحات مفصلة لكتاب معين، ولا يشرح كل مميزات برنامج ستاتا. فهذا الكتاب يوضح كيف يمكن استخدام برنامج ستاتا لإنجاز مجموعة كبيرة من المهام الإحصائية.

فصول هذا الكتاب تغطي موضوعات عامة بدلاً من التركيز على أوامر ستاتا معينة، وهذا يعطي الكتاب شكلاً مختلفاً عن أدلة استخدام برنامج ستاتا. فمثلاً فصل إدارة البيانات يغطي العديد من الطرق المختلفة لإنشاء أواستيراد أودمج أو إعادة ترتيب ملفات البيانات. والفصول الخاصة بالرسومات البيانية، والإحصائيات الملخصة، والجداول، وتحليل التباين، ونماذج المقارنة الأخرى، لها نفس هذه الخطوط العريضة التي تشمل عددًا من التقنيات المختلفة. وهناك فصل جديد يعطي مقدمة عن بيانات الدراسات الاستقصائية، والذي تم وضعه ضمن الفصول الأولى من هذا الكتاب، حيث يعطي خلفية مع أمثلة عن ذلك في الأماكن المناسبة والفصول اللاحقة.

الموضوعات العامة لأول سبعة فصول (من بداية تحليل الانحدار الخطّي) تتناسب مع مادة الإحصاء التطبيقي للدراسات الجامعية، أو المستويات الأولى من الدراسات الجامعية، ولكن تم إضافة عمق إضافي لتغطية نضايا خاصة عادةً يواجهها المحللون، مثل كيفية استيراد البيانات أو الرسومات البيانية لغرض النشر العلمي، أو جودة الأشكال البيانية، أو العمل مع أوزان الدراسات الاستقصائية، وحل مشاكل الانحدار. أما الفصل "8" (طرق انحدار متقدمة) والفصول اللاحقة له، فقد قمنا بالانتقال إلى موضوعات متقدمة، أو ما يُسمى البحث الواقعي، حيث يمكن للقارئ أن يجد

معلومات أساسية وشروحات عن دالة تمهيد لوس، أو الانحدار الموثوق، أو الانحدار الربيعي، أو الانحدار غير الخطي، أو الانحـــدار اللوغـــاريتمي، أو الانحدار اللوغاريتمي المرتب، أو الانحدار اللوغاريتمي متعدد الحـــدود، أو انحدار بواسون. وتم تطبيق طرق جديدة لـــصياغة المعادلـــة الهيكليـــة، أو الحساب المتعدد للقيم المفقودة، ونماذج حساب الأحداث، وزمن البقاء. كما تم أيضاً صياغة واستخدام المتغيرات المركبة من التحليل العاملي، أو المكونات الرئيسة، حيث تم تقسيم المشاهدات إلى أنواع تجريبية، أوعناقيد، أو تحليـــل سلاسل زمنية متعددة أو بسيطة لتناسب نماذج التأثيرات المختلطة أو متعددة المستويات. كما قامت شركة ستاتا بالعمل بجد في السنوات الأخيرة لتطـــوّر مكانتها الفنية والتقنية، ويظهر هذا المجهود بوضوح في تعدد أنواع أوامـــر النماذج الإحصائية التي يستخدمها البرنامج الآن.

كما أن الكتاب يلقي نظرة على البرمجة في ستاتا. حيث يجد العديد من القرّاء بأن برنامج ستاتا يقوم بكل شيء يحتاجونه، ولذلك فليس هنـــاك داع لكتابة البرامج الأصلية، بالنسبة للبعض، فإن البرمجة في ستاتا تعتبر إحـــدى الميزات الجذابة التي أدت إلى التطور السريع لهذا البرنامج.

أما الفصل (14) فيفتح الباب للمستخدمين الجدد لاكتشاف البرمجة فـــي ستاتا، وكيفية استخدام هذه البرمجة لأغراض مهام الإدارة المتقدمة للبيانـــات أو غيرها، حيث يمكن للمستخدمين إنشاء قدرات إحصائية جديـــدة لتجـــارب مونت كارلو أو لغرض التدريس.

وعموماً، هناك إصدارات متشابهة "بتنسيقات مختلفة" لبرنامج ستاتا، مع أنظمة تشغيل ويندوز وماك ويونيكس. وفي كل أنظمة التشغيل هـــذه، فـــإن برنامج ستاتا يقوم باستخدام نفس الأوامر، ويقوم بإنتاج نفس المخرجات، كما أن البيانات، والأشكال البيانية، والبرامج التي يتم إنشاؤها باستخدام ستاتا مع نظام تشغيل معين يمكن استخدامها مع نظام تشغيل آخر. التنسيقات المختلفة لبرنامج ستاتا – التي تظهر باختلاف نظام التشغيل – عبارة عـــن اخـــتلاف

بسيط في تفاصيل الشاشة، والقوائم والتعامل مع الملفات، حيث يتبع برنـــامج ستاتا ما هو متعارف عليه في كل نظـــام التـــشغيل، مثـــل تحديـــد المجلـــد \directory\filename في ظل تشغيل ويندوز، وعلى العكس من ذلـــك، فـــإن المجلد في ظل نظام يونيكس يتم تحديده directory/filename/؛ وبـــدلاً مـــن استخدام عرض لأنظمة التشغيل الثلاثة، فقد قمت باســـتخدام نظـــام تـــشغيل ويندوز. وبالنسبة لمستخدمي أنظمة التشغيل الأخرى، فسوف يحتاجون للقيام بعمليات تحويل بسيطة إذا احتاجوا لذلك.

المؤلف



ملاحظات المؤلف على الطبعة الثامنة

بدأتُ في استخدام برنامج ستاتا سنة 1985، وهي أول سنة يتم فيها إطلاق هذا البرنامج. ومبدئياً فإن برنامج ستاتا كان يعمل على نظام تشغيل MS-DOS فقط، ولكن الاتجاه نحو استخدام إصدار مع أجهزة الكمبيوتر الشخصي، جعل البرنامج يبدو أكثر حداثة من منافسيه الذين ظهروا قبل ثورة الكمبيوتر. حيث كانت هناك كروت مثقبة كبيرة مع 80 عمودًا تعمل في بيئة الفورتران. وعلى خلاف الاتجاه العام في البرامج الإحصائية التي تعتبر كل مستخدم عبارة عن حزمة من الكروت، فإن برنامج ستاتا نظر إلى التعامل مع المستخدِم كنوع من المحادثة، فطبيعة تفاعله وتكامله مع العمليات الإحصائية لإدارة البيانات، والأشكال البيانية تدعم التدفق الطبيعي للتفكير التحليلي بعدة طرق، لم تستطع البرامج الأخرى القيام بها. فالأمر **graph**، والأمر **predict**، أصبحا من الأوامر المفضلة لدى الكثير، لقد أعجبني كثيراً كيف أن كلها تتناسب معاً. وبدأتُ كتابة أول كتاب عن ستاتا بعنوان "*الإحصاء مع ستاتا*" والذي نُشِرَ سنة 1989 لإصدار ستاتا 2. والذكرى العشرون لبرنامج ستاتا في سنة 2005 كانت متميزة بنشر إصدار خاص من مجلة ستاتا *Stata Journal* يحتوي على مجموعة مقالات تاريخية، ومقابلات، وملخص تاريخي عن كتاب "الإحصاء مع ستاتا".

ستاتا تغير بقدر كبير جداً منذ صدور الطبعة الأول للكتاب، حيث لاحظت أن ستاتا ليس برنامجاً للقيام بكل شيء، ولكن الأشياء التي يقوم بها يتم إنجازها بشكل رائع. التوسع الكبير في قدرات ستاتا كان مذهلاً، وهذا واضح جداً في التكاثر، ولاحقاً في التبرير المنطقي لإجراءات صياغة النماذج. فأسلوب بناء وليام جولد لبرنامج ستاتا، والذي يشمل أدوات ستاتا

البرمجية وصيغ أوامره نصيّة تم تقديمها بشكل جيد، وتم دمج طرق إحصائية جديدة خلال تطويرها، وهناك تشكيلة كبيرة من الرسومات البيانية في الفصل (3) وفي الفصل (8) حيث تمت مناقشة عدد كبير من أوامر صياغة النماذج.

وفي الفصول اللاحقة تم مناقشة السلاسل الزمنية الجديدة، وقدرات النماذج المختلطة، ونماذج التعويض المتعددة. وهذا النقاش يوضح مدى التطور الكبير في برنامج ستاتا خلال السنوات الماضية، كما توجد هناك مساحات للتقنيات الجديدة مثل تلك المتعلقة بالبيانات الطولية (**xt**)، والبيانات الاستقصائية (**svy**)، وبيانات السلاسل الزمنية (**ts**)، وزمن البقاء (**st**)، والإسناد المتعدد (**mi**). هذه المساحات تفتح عالم الاحتمالات، كما تفعل الأوامر البرمجية للنماذج الخطية العامة (**glm**)، أو الإجراءات العامة لتقدير الأرجحية العظمى؛ التطور الرئيس الآخر يتضمن تطور قدرات برمجة المصفوفات، والزيادة بإضافة مميزات جديدة في إدارة البيانات والأدوات التحليلية متعددة الأغراض، مثل الرسومات البيانية الهامشية أو صياغة المعادلة الهيكلية، وإدارة البيانات تم تطويرها من موضوع بسيط في أول طبعة لكتاب *الإحصاء مع ستاتا* إلى فصل كامل في الطبعة الثامنة.

القائمة الشاملة لبرنامج ستاتا، ونظام مربعات الحوار، تعتبر طريقة بديلة للأوامر التي تتم طباعتها. حيث يمكن الاختيار والنقر على أي أمر، القوائم المنسدلة واختيارات ومربعات الحوار يمكن تعلمها بسهولة من خلال الاكتشاف بدلاً من القراءة. وعموماً فإن كتاب "*الإحصاء مع ستاتا*" يعطي تلميحات عامة عن القوائم في بداية كل فصل. في أغلب أجزاء هذا الكتاب، قمنا باستخدام الأوامر، وذلك لتوضيح ما يمكن لبرنامج ستاتا عمله، ونظائر الأوامر الموجودة بالقوائم يمكن التعرّف عليها بسهولة، وعلى العكس من ذلك، فإذا بدأت العمل من خلال القوائم، فإن ستاتا يوفر تدريباً غير رسمي من خلال عرض الأمر ذات العلاقة في نافذة النتائج، حيث إن القوائم ومربعات الحوار تعمل على ترجمة النقر إلى أوامر ستاتا، والتي يتم تنفيذها ببرنامج ستاتا.

الرسومات البيانية التحليلية تعتبر إحدى نقاط القوة المتميـزة ببرنـامج ستاتا، كما هو واضح في كل فصل من هذا الكتاب. فالعديد من الأمثلة ليست صورًا بدائية تقوم بتوضيح تقنية معينة، ولكنها تقوم بتجسيد بعض التحسينات لغرض النشر أو تطوير جودة العرض. قد يقوم القارئ بتصفح هذه الأشكال البيانية للتعرف على بعض الأفكار حول الأشكال البيانية المحتملة، والتي قد لا تعرضها أدلة استخدام برنامج ستاتا.

الإحصاء مع ستاتا (الإصدار 12) يختلف جوهرياً عن سابقه – وهـو كتاب (الإصدار 10) حيث تم إعادة تنظيم الفصول، وتم تضمين فصل جديد وهو عبارة عن مقدمة لبيانات الدراسات الاستقصائية، والذي يأتي في بدايات الكتاب؛ كما تم تناول مواضيع الانحدار في أربعة فصول في كتاب الإصدار 10 تم دمجها وتنظيمها بطريقة أكثر منطقية في فصلين، الأول: عن الانحدار الخطي، أما الثاني: عن طرق الانحدار المتقدمة. فصل الانحـدار المتقـدم يحتوي على أجزاء جديدة عن الإسناد المتعدد للقيم المفقودة، وعـن نمـاذج المعادلة الهيكلية (SEM)، أما فصل تحليل المكونات الأساسـية، والتحليـل العاملي، والتحليل العنقودي، فيتضمن أيضاً جزءين جديدين يعرضان استخدام العلامات العامليّة في الانحدار، واستخدام قياس النماذج فـي SEM، الجـزء الجديد في فصل صياغة نمـاذج التـأثيرات المختلطـة، وذات المـستويات المتعددة، يعرض تجربة القياس المتكرر، أما الفصل الأخيـر، فهـو عـن البرمجة، وتم تبسيطه وتركيزه حول مثال رئيس (يقوم برسم أشكال بيانيـة للمسح المتعدد)، وهو مفيد لبعض القرّاء.

أحد أهداف إصدار ومراجعة كتاب الإصدار 12، هو تطـوير الأمثلـة فبعضها كان من بحوثي من فترة التسعينيات، ولكن مازالت مفيدة، فتحليـل مكوك الفضاء تشالينجر – والذي تم استخدامه في إصدار سنة 1989 من هذا الكتاب والإصدارات اللاحقة – مازال مفيدًا لعرض الأفكار الأساسـية فـي بداية فصل الانحدار اللوغاريتمي، هذا الفصل ينتهي مع التحليل اللوغاريتمي متعدد الحدود للردود على استطلاع سنة 2011، الذي يقوم بالاستفسار عـن ماذا يعرف الناس؟ وماذا يعتقدون بخصوص التغير المنـاخي، الاسـتطلاع

الخاص بالتغير المناخي هو واحد من ثلاثة بيانات استطلاعية جديدة لسنة 2010 وسنة 2011. وهذه البيانات تم استخدامها في عدة أمثلة في فصول مختلفة، أحد هذه الفصول (تحليل المكونات الرئيسة والتحليل العاملي) يبدأ مع بيانات مبسطة خاصة بالكواكب، ولكن ينتهي بجزء جديد يدمج التحليل العاملي مع الانحدار، وقياسات مناظرة، ونماذج المعادلة الهيكلية مستخدماً بيانات استطلاع عن البيئة الساحلية لسنة 2011؛ الأمثلة الأخرى تتضمن سلاسل زمنية للمتغيرات التنبؤية للمناخ نفسه. وأحد البيانات المتميزة تتعلق بـ 42 قرية في آلاسكا، وتم الحصول على هذه البيانات من دراسة في سنة 2011، وهي توضح كيف أن صياغة نماذج التأثيرات المختلطة يمكن أن تدمج العلوم التطبيقية مع بيانات العلوم الاجتماعية. ونماذج ARMAX تختتم فصل السلاسل الزمنية والتي تم تطويرها ببيانات دراسة أجريت في سنة 2011 اختبرت "الإشارة الحقيقية" للاحتباس الحراري، وحيثما أمكن فقد حاولت أن تكون الأمثلة تطرح أسئلة بحثية حول قضايا عامة بدلاً من عرض أرقام لتوضح طريقة التحليل. كما أن العديد من بيانات الأمثلة تتضمن متغيرات أخرى تتجاوز ما تم مناقشته في هذا الكتاب، وهي تمثل دعوة للقرّاء بأن يقوموا بتحليلات أكثر خاصة بهم.

وكما لاحظنا في الفصل (1)، فإن أدوات المساعدة والبحث في ستاتا تم تطويرها لتحافظ على السرعة مع البرنامج، وبالإضافة إلى وثائق ستاتا المتوافرة من خلال ملفات المساعدة، توجد هناك صفحة إنترنت خاصة ببرنامج ستاتا، وإمكانية البحث في الوثائق الموجودة بتلك الصفحة. كما يوجد هناك منتدى المستخدمين، وتوجد هناك برامج تدريبية في NetCourses، ومجلة ستاتا *Stata Journal*. وهناك أكثر من 9,000 صفحة من الوثائق. فكتاب "*الإحصاء مع ستاتا*" يعتبر بوابة عبور لبرنامج ستاتا. فكل هذه المصادر التي تم ذكرها سوف تكون مصادر لمساعدتك.

مقدمة المترجم

الحمد لله، والصلاة والسلام على رسول الله، وعلى آله وصحبه أجمعين. الحمد لله الذي وفقني لإتمام ترجمة هذا الكتاب، فلله الحمد والمنّة على ذلك. هذا الكتاب هو محاولة لسد النقص الملحوظ في المراجع العربية التي تتناول بالشرح والتحليل كيفية استخدام برنامج ستاتا في التحليل الإحصائي. فهذا البرنامج له قدرات هائلة تخفى عن الكثيرين. وهذا الكتاب هو بمثابة دليل استخدام مبدئي لبرنامج ستاتا. فمن الصعب الإلمام بكافة قدرات هذا البرنامج وإمكاناته الضخمة في كتاب واحد.

قبل الولوج إلى دفات كتاب "**الإحصاء مع برنامج ستاتا**" أنصح القارئ الكريم بأن يتجه مباشرة للموضوعات ذات الصلة بدراسته واهتماماته، مع ملاحظة ضرورة الاطلاع على الفصول الثلاثة الأولى؛ حيث تناول **الفصل الأول:** لمحة سريعة عن الكتاب، وما تعنيه تنسيقات الخطوط المستخدمة. أما **الفصل الثاني:** فقد تطرق إلى كيفية إدخال البيانات لبرنامج ستاتا، وكيفية استيرادها من البرامج الأخرى، وإدخالها لبرنامج ستاتا، وكذلك نوّه إلى أنواع المتغيرات، وأنواع العينات. أما **الفصل الثالث:** فهو بمثابة معرّض لأشكال الرسومات البيانية، حيث شرح بنوع من التفصيل أوامر إنشاء الرسومات البيانية وأنواعها وطرق تحريرها وتعديلها، كما تناول كذلك كيفية دمج أكثر من شكل بياني في شكل واحد.

بقية فصول الكتاب – وهي عشرة فصول – فقد تناولت موضوعات تحليلية مستقلة بذاتها، فيمكن للقارئ الانتقال إلى الفصل الذي يتعلق بمجال بحثه، وفي حالة الحاجة إلى أي معلومة من الفصول السابقة، فقد تم الإشارة إلى الجزء الذي يجب على القارئ الانتقال إليه مباشرة في الفصول السابقة؛

أما الفصل الأخير – وهو **الفصل الرابع عشر** – فيُعتبر مقدمة عن كيفية إنشاء أوامر ببرنامج ستاتا. وهذه الميزة لا توجد في العديد من برامج الحزم الإحصائية المعروفة، حيث يمكن للقارئ تصميم أوامر لأساليب إحصائية جديدة – وهذا الأمر متقدم جداً – أو إدارة قاعدة بيانات معينة، وجعل هذه الأوامر متاحة لجميع المستخدمين في العالم ليقوموا بتحميلها واستخدامها.

ختاماً، لا يفوتني أن أتقدم بخالص شكري وامتناني لكن من ساهم في إخراج هذا الكتاب إلى حيّز الوجود، وأخص بالذكر والديّ فالكلمات تعجز عن التعبير عن مدى امتناني لهما، كما أتقدم بشكري الجزيل لإخوتي وأصدقائي الأعزاء على تشجيعهم ودعمهم المستمر على إتمام هذا العمل، كما أتقدم بشكري وامتناني لأساتذتي الأفاضل، وأخص بالذكر أستاذي الفاضل وصديقي العزيز الدكتور إبراهيم أحمد بالخير، فهو مثال للعطاء وقدوة في البذل بإخلاص، لقد كان لدعمه وتشجيعه لي خلال فترة دراستي بمرحلتي الماجستير والدكتوراه الأثر الكبير في تطوير مهاراتي. كما أتقدم بالشكر لجميع العاملين بدار المريخ على مجهوداتهم وتعاونهم الكبير في إخراج هذا الكتاب بالصورة المطلوبة. فجزى الله الجميع عني خيراً، وأدعو الله أن يجعل هذا الكتاب علماً يُنتفَعُ به.

د. رمضان مفتاح الفيتوري

كلية الاقتصاد – جامعة عمر المختار – ليبيا

تنويه:

لتطبيق الأمثلة الواردة في هذا الكتاب، فأنت تحتاج إلى تحميل ملفات البيانات الخاصة بالأمثلة، ويمكنك الحصول عليها من الرابط التالي: http://www.stata.com/bookstore/swsdl.html

المترجم

الفصل الأول

ستاتا ومصادر البيانات

Stata and Stata Resources

ستاتا عبارة عن برنامج إحصائي متكامل لأجهزة الكمبيوتر التي تعمل باستخدام نظم التشغيل Windows أو Mac أو Linux، حيث إنه يمتاز بالسهولة والسرعة في الاستخدام، وهو عبارة عن مكتبة لها القدرة على إدارة البيانات واستخدام البرامج التحليلية المعدة مسبقاً، والقدرة على البرمجة التي تتيح للمستخدمين اختراع وإضافة قدرات أكثر حسب الحاجة. أغلب العمليات الإحصائية يمكن إنجازها باستخدام القوائم المنسدلة أو بطباعة الأوامر مباشرة. كما أن القوائم تساعد المستخدمين الجدد لبرنامج ستاتا على تعلّم البرنامج، كما أنها تساعد على تطبيق الإجراءات غير المعتادة؛ إن استخدام أوامر ستاتا وكتابتها باستمرارتساعد المستخدمين ذوي الخبرة على القيام بأعمالهم بكفاءة أكثر، كما يجعل تطوير البرامج للاستخدامات المعقدة عملية سهلة. أما استخدام القوائم والأوامر معاً، يمكن أن يتم حسب الحاجة أثناء استخدام ستاتا. كما أن المساعدة المكثّفة الموجودة ببرنامج ستاتا، تمكنك من البحث ومعرفة الميزات التي تجعل من السهل العثور على تركيبة الأمر والمعلومات في وقت قصير جداً. هذا الكتاب تم إعداده كتكملة لميزات المساعدة الموجودة ببرنامج ستاتا.

بعد مقدمة مبسطة، سوف نبدأ بأمثلة عن ستاتا لإعطائك لمحة عن تسلسل عملية تحليل البيانات، وكيفية استخدام النتائج الإحصائية. والفصول اللاحقة تشرح ذلك بشكل أكثر تفصيلاً. حتى بدون شرح، يمكنك أن ترى كيف أن أوامر ستاتا واضحة، فمثلاً **use *filename*** هو أمر يُستخدم لفتح ملف بيانات اسمه filename، أما الأمر **summarize** فيمكنك عن طريقه

الحصول على ملخص إحصائي، **correlate** للحصول على مصفوفة الارتباط وهكذا، هذه النتائج أيضاً يمكن الحصول عليها باختيارها من قوائم Data أو Statistics.

مستخدمو ستاتا لديهم عدد كبير من المصادر التي تساعدهم على تعلّم ستاتا وحل المشاكل مهما كانت صعبة. هذه المصادر تأتي ليس فقط من شركة ستاتا نفسها، بل أيضاً من عدد من المستخدمين النشطين. وسوف تعرض أجزاء هذا الفصل المصادر الرئيسة للحصول على المساعدة وطباعة دليل الاستخدام، وأين يمكنك إرسال بريد إلكتروني للحصول على المساعدة التقنية من صفحة ستاتا www.stata.com والتي توفر الكثير من الخدمات بما فيها التحديثات، وجزء عن الأسئلة المتكررة، وقائمة ستاتا على الإنترنت، ومجلة ستاتا المحكّمة.

ملاحظات حول تنسيقات الطباعة : A Typographical Note

هذا الكتاب يستخدم مجموعة من التنسيقات التي توضح كيفية استخدام الكلمات:

- الأوامر التي تُطبع بواسطة المستخدم تظهر بخط أسود عريض، وعند كتابة سطر أمر بالكامل، فإنه يبدأ بمسافة، كما يظهر في نتائج ستاتا أو ملف سجل المخرجات.

.correlate extent ***area volume temp***

- أسماء المتغيرات والملفات الموجودة في أي أمر تظهر بخط مائل للتأكيد على أنها ليست جزءًا من الأمر ولكنها تابعة له.
- أسماء المتغيرات والملفات تظهر بخط مائل ضمن الأمر نفسه حتى تساعد على التفرقة بينها وبين الكلمات العادية.
- العناصر التي تُعبّر عن قوائم من شريط مهام ستاتا تظهر بخط من نوع Arial يليها الخيارات ويفصل بينها علامة ">". فعلى سبيل المثال، يمكننا فتح ملف بيانات باستخدام القائمة File > Open، وبعد ذلك يتم تحديد

الملف المطلوب والنقر عليه؛ وهناك العديد من المهام التي يمكن إنجازها باستخدام القوائم الموجودة في أعلى نافذة ستاتا.

File Edit Data Graphics Statistics User Window Help

أو استخدام صف الأيقونات الموجودة أسفل هذه القــوائم، فمــثلاً اختيار القائمة File > Open يُعادِل النقر على الأيقونة الموجــودة أقــصى اليسار، وهي تشبه المجلد المفتوح، نفس الشيء يمكن القيام به بطباعة أمــر على الشكل التالي:

```
.use filename
```

وهكذا يمكننا عرض ملخص إحصائي للمتغير *extent* كما يلي :

```
.summarize extent
```

Variable	Obs	Mean	Std. Dev.	Min	Max
extent	33	6.51697	.9691796	4.3	7.88

هذه التنسيقات توجد فقط في هذا الكتاب، ولا توجد في برنامج ســتاتا نفسه، وستاتا له القدرة على عرض أنواع متعددة من الخطوط على الشاشة، ولكنه لا يستخدم الخط المائل في أوامر ستاتا، فعندما يقوم برنــامج ســتاتا باستخدام ملفات معينة تم استيرادها من برنامج نصي آخر أو جدول نتائج تم نسخه ولصقه، يمكنك تغيير تنسيق هذه الملفات إلى نوع الخط Courier وحجم 10 أو أصغر، وبذلك فإن أعمدة البيانات سوف تظهــر بــصورة منظمــة وواضحة.

ستاتا برنامج حساس لحالة أحرف المتغيرات، حيث إن الحرف الصغير يختلف عن الحرف الكبير، ولذا فــإن **summarize** يعتبــر أمــرًا ولكــن Summarize وSUMMARIZE لا تعتبر أوامر، كذلك فإن *Extent* و*extent* ســوف يتم اعتبارهما متغيرين مختلفين.

مثال على استخدام ستاتا : An Example Stata Session

كنظرة عامة على كيفية استخدام ستاتا، هذا الجزء سوف يشرح استخدام وتحليل بيانات تم إعدادها مسبقاً بملف اسمه *Arctic9.dta**، وهي عبارة عن سلسلة زمنية صغيرة تم جمعها بالأقمار الصناعية في الفترة (1979 إلى 2011) فهي مشاهدات عن الجليد في المحيط المتجمد الشمالي خلال شهر سبتمبر في أقل نقطة من حلقته السنوية. البيانات تم الحصول عليها من ثلاثة مصادر مختلفة (انظر المرفق الخاص بمصادر البيانات في نهاية هذا الكتاب). المتغير الأول *extent* وهو قياس يعتمد على صور القمر الصناعي للمنطقة البحرية في نصف الكرة الشمالي والتي تحتوي على نسبة 15% على الأقل من الجليد كل شهر سبتمبر *Area*. وهي أرقام أقل من *extent* تمثل المنطقة المتجمدة نفسها، متغير آخر وهو *tempN* يوضح متوسط درجة الحرارة السنوية لهواء سطح البحر بالمنطقة التي تقع أعلى من °64 شمال خط الاستواء، ودرجات الحرارة تم التعبير عنها بدرجات الحرارة الشاذة، وهي تلك المكتوبة بطريقة مختلفة، بحيث إنها توضح الانحراف عن متوسط درجة الحرارة المئوية خلال الفترة 1951 – 1980م؛ حيث لدينا 33 مشاهدة (سنوات) و8 متغيرات.

نحن ربما نحتاج في نهاية الأمر إلى تسجيل خطوات هذه العملية، وأفضل طريقة للقيام بذلك تتم بواسطة فتح ملف تسجيل في بداية عملنا، حيث إن ملف التسجيل يحتوي على الأوامر، وجداول النتائج، ولكن لا يحتوي على الرسومات البيانية. وللبدء بفتح ملف تسجيل قم بختيار File > Log > Begin من شريط القوائم، ثم قم باختيار اسم الملف والمجلد الذي تريد أن تحفظ فيه ملف تسجيل النتائج، أو يمكنك طباعة أمر مباشر كما يلي :

```
.log using Monday1
```

هناك عدة طرق للقيام بالأشياء الأكثر استخداماً في ستاتا، وكل طريقة لها مزاياها وتناسب أوضاعاً أو أذواقاً مختلفة للمستخدمين.

* البيانات الخاصة بهذا الكتاب يمكن الحصول عليها من موقع www.stata.com

ملفات التسجيل يمكن إنشاؤها في ستاتا بامتداد خاص (.smcl)، أو كملفات نصية عادية أو بتنسيق ASCII (.log)، امتداد smcl. (لغة ستاتا في الاستعادة والتحكم) سوف تكون مناسبة للمعاينة والطباعة باستخدام برنامج ستاتا، كما يمكن أيضًا أن يحتوي الملف على اختصار يساعد في فهم الأوامر ورسائل الخطأ، والتي لا يمكن أن تحتويها ملفات التسجيل النصية، ولكن الملفات النصية لها نفس الاستخدامات إذا كنت تريد التخطيط لاحقاً لإدراج أو تحرير نتائج في برنامج معالج النصوص Word؛ بعد اختيار نوع ملف التسجيل تحتاج إلى حفظ الملف بواسطة الضغط على Save. في هذا التمرين، سوف نقوم باستخدام ملف تسجيل بنوع smcl باسم *Monday1.smcl*.

سوف يتم تحليل بيانات الملف المسمى *Arctic9.dta*. ولفتح الملف واستخراج بياناته، هناك عدة خيارات، من قائمة المهام اختر منها File > Open >*Arctic9.dta*

أو اضغط على أيقونة فتح ثم *Arctic9.dta*

أو قم بطباعة الأمر:

```
.use Arctic9.dta
```

الإعدادات الافتراضية في نظام التشغيل ويندوز تجعل ستاتا يبحث عن ملفات البيانات في مجلد المستندات الخاصة بالمستخدم، وإذا كنا نريد ملفاً في مجلد مختلف، فإنه يجب علينا تحديد موقع الملف عند استخدام أمر **use**

```
.use C:\books\sws_12\data\Arctic9
```

أو يمكنك تغيير المجلد الافتراضي وذلك باستخدام الأمر **cd**

```
.cd C:\books\sws_12\data\
.use Arctic9
```

أو قم باختيار ... File >Change Working Directory من القوائم، وفي الغالب فإن أسهل طريقة لفتح ملف البيانات هي File > Open ثم تحديد المجلد بالطريقة المعتادة.

لمشاهدة وصف مختصر للبيانات قم بطباعة الأمر:

```
.describe
```

```
Contains data from C:\data\Arctic9.dta
  obs:            33                          Arctic September mean sea ice 1979-2011
 vars:             8                          2 Jul 2012 06:11
 size:           891

              storage  display    value
variable name   type   format     label      variable label

year            int    %ty                   Year
month           byte   %8.0g                 Month
extent          float  %9.0g                 Sea ice extent, million km^2
area            float  %9.0g                 Sea ice area, million km^2
volume          float  %8.0g                 Sea ice volume, 1000 km^3
volumehi        float  %9.0g                 Volume + 1.35 (uncertainty)
volumelo        float  %9.0g                 Volume - 1.35 (uncertainty)
tempN           float  %9.0g                 Annual air temp anomaly 64N-90N C

Sorted by:  year
```

العديد من أوامر ستاتا يمكن اختصارها باستخدام حروفها الأولى، فمثلاً يمكننا اختصار **describe** لتكون فقط الحرف **d**، باستخدام القوائم يمكن الحصول على نفس الجدول بختيار الأمر:

Data>Describe Data > Describe data in memory > (OK).

البيانات الموجودة لدينا تحتوي على 33 مشاهدة و8 متغيرات. لذا يمكننا وضع قائمة بكل محتوياتها بطباعة الأمر **list** (أو الحرف **l**، أو اختيار (Data>Describe Data > List data > (OK)، للاختصار فقط سوف نعرض أول عشر سنوات بطباعة الأمر:

.list in 1/10

	year	month	extent	area	volume	volumehi	volumelo	tempN
1.	1979	9	7.2	5.72	16.9095	18.2595	15.5595	-.57
2.	1980	9	7.85	6.02	16.3194	17.66937	14.96937	.33
3.	1981	9	7.25	5.57	12.8131	14.16307	11.46307	1.21
4.	1982	9	7.45	5.57	13.5099	14.85987	12.15987	-.34
5.	1983	9	7.52	5.83	15.2013	16.5513	13.8513	.27
6.	1984	9	7.17	5.24	14.6336	15.98357	13.28357	.31
7.	1985	9	6.93	5.36	14.5836	15.93363	13.23363	.3
8.	1986	9	7.54	5.85	16.0803	17.43027	14.73027	-.05
9.	1987	9	7.48	5.91	15.3609	16.7109	14.0109	-.25
10.	1988	9	7.49	5.62	14.988	16.338	13.638	.87

التحليل يمكن أن يبدأ بجدول به المتوسط الحسابي والانحراف المعياري، وأعلى قيمة وأصغر قيمة، بطباعة الأمر **summarize** أو **su**، أو قم باختيار ذلك من القائمة

Statistics>Summaries, tables, and tests > Summary and descriptive statistics > Summary statistics (OK)

```
.summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
year	33	1995	9.66954	1979	2011
month	33	9	0	9	9
extent	33	6.51697	.9691796	4.3	7.88
area	33	4.850303	.8468452	3.09	6.02
volume	33	12.04664	3.346079	4.210367	16.9095
volumehi	33	13.39664	3.346079	5.560367	18.2595
volumelo	33	10.69664	3.346079	2.860367	15.5595
tempN	33	.790303	.7157928	-.57	2.22

لطباعة نتائج هذا التحليل اضغط على نافذة النتائج ثم اضغط أو من القوائم اختر File>Print > Results

لنسخ الجداول والأوامر أو أي معلومات أخرى من نافذة النتائج في أي معالج نصوص، قم باستخدام الفارة لتحديد النتائج المراد نسخها ثم اضغط الزر الأيمن للفارة واختر Copy Text من القائمة المختصرة، ثم بعد ذلك انتقل إلى معالج النصوص، واختر المكان المراد إدراج البيانات به، ثم اضغط على الزر الأيمن للفارة واختر Paste أو اضغط على أيقونة لصق في برنامج معالج النصوص Microsoft Word؛ الخطوة الأخيرة في أغلب الحالات هي تعديل النص الذي تم لصقه ليكون بحجم الخط المطلوب.

مستوى جليد البحر *extent* وحجمه *volume* في شمال الكرة الأرضية *area* يفترض أن يرتبط بدرجة حرارة الهواء السنوية *tempN*، ليس فقط لأن الهواء الساخن يساعد في الذوبان، ولكن أيضاً لأن درجة الهواء فوق سطح الجليد سوف تكون أكثر دفئاً من الجليد نفسه؛ يمكننا مشاهدة الارتباط بين هذه المتغيرات بطباعة الأمر **correlate** يليه قائمة بالمتغيرات

```
.correlate extent area volume tempN
```

```
(obs=33)

             |   extent     area   volume    tempN
-------------+------------------------------------
      extent |   1.0000
        area |   0.9826   1.0000
      volume |   0.9308   0.9450   1.0000
       tempN |  -0.8045  -0.8180  -0.8651   1.0000
```

النتائج توضح أن هناك ارتباطاً إيجابياً قوياً بين مستوى جليد البحر *extent* وحجمه *volume* في شمال الكرة الأرضية *area* وهو ما كان متوقعاً، وهذه المتغيرات ترتبط بشكل سلبي مع درجة الحرارة *tempN*، وهذا يعني أن درجة حرارة الهواء تؤدي إلى نقص في الجليد والعكس صحيح.

مصفوفة الارتباط هذه، يمكن الحصول عليها كذلك باستخدام القوائم كما يلي:

Statistics>Summaries, tables, and tests > Summary and descriptive statistics >Correlation and covariance

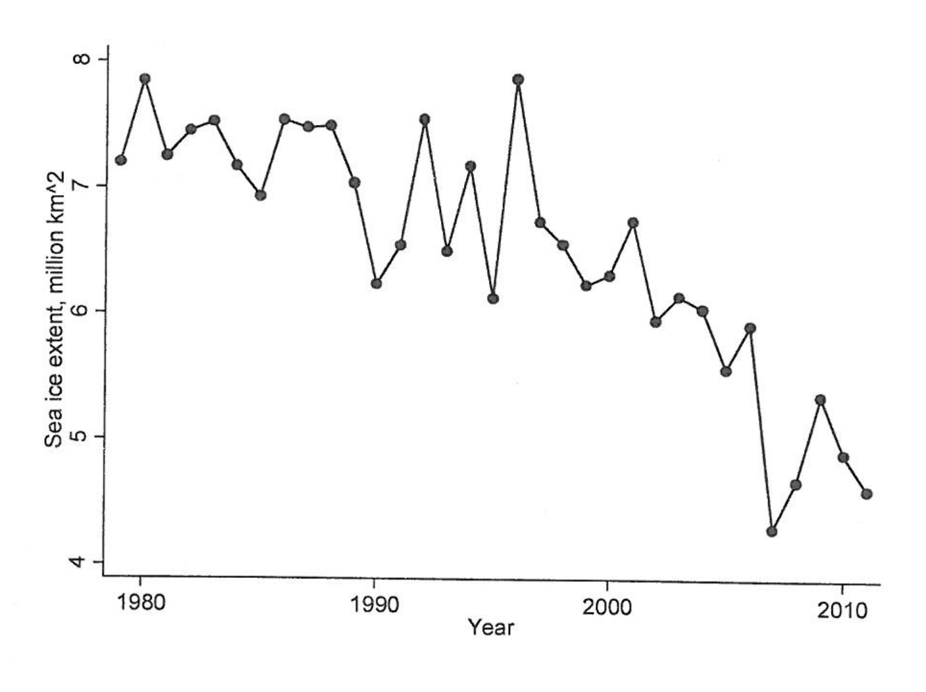
ثم قم باختيار المتغيرات التي تريد قياس ارتباطها.

بالرغم من أن استخدام القوائم سهل وواضح، فإنها أكثر تعقيداً من طباعة الأوامر، وبناءً على ذلك، فإننا سوف نركز بشكل أساسي على طباعة الأوامر، وفي نفس الوقت سوف نشير إلى القوائم بشكل عرضي. أما اكتشاف كيفية استخدام القوائم، وكيفية القيام باستخراج النتائج عن طريق هذه القوائم فسوف نتركه للقارئ. والسبب الآخر في استخدام طريقة طباعة الأوامر، هو أن دليل استخدام ستاتا يستخدم طريقة الأوامر للشرح والتوضيح.

إذن فهناك ارتباط بين مستوى جليد البحر *extent* وحجمه *volume* في شمال الكرة الأرضية *area* ودرجة الحرارة *tempN*، كيف تأثرت هذه المتغيرات خلال فترة من الزمن؟ الرسم البياني (1.1) والذي تم الحصول عليه بالأمر **graph twoway connect** يوضح مستوى الجليد خلال سنوات الدراسة *year*، المتغير الذي يتم إدخاله أولاً في هذا الأمر هو مستوى جليد

البحر *extent*، سوف يظهر على المحور العمودي أو محور y بينما المتغير الذي يتم إدخاله أخيراً فسوف يظهر على المحور الأفقي year أو محور *x*، ويمكننا أن نرى من الشكل أن هناك انخفاضاً كبيرًا في المنحنى، حيث إن حجم الجليد خلال شهر سبتمبر انخفض بأكثر من الثلث خلال المدة.

```
.graph twoway connect extentyear
```



الشكل (1.1)

لطباعة الرسم البياني، اذهب إلى نافذة الرسم البياني، واضغط على أيقونة الطباعة أو اختر File < Print، ولنسخ الرسم البياني داخل معالج النصوص اضغط زر الفارة الأيمن على الرسم البياني ثم اختر Copy بعد ذلك انتقل إلى معالج النصوص وحدد المكان الذي تريد أن تقوم بإدراج الرسم به واضغط أيقونة لصق أو اختر تحرير > لصق، أو تحرير > لصق خاص.

لحفظ الرسم البياني للاستخدام مستقبلاً اضغط زر الفارة الأيمن واختر Save as أو اضغط أيقونة حفظ في نافذة الرسم البياني، أو اختر < File

Save ومن قائمة Save as type هناك مجموعة تنسيقات مختلفة لنوع الملف، في نافذة الحفظ هناك عدة خيارات بامتدادات مختلفة، وهذه الخيارات تتضمن:

Stata graph (*.gph) وهذا النوع يحتوي على معلومات كافية ليتم تحرير الرسم باستخدام برنامج ستاتا.

As-is graph (*.gph) وهو نوع مضغوط من تنسيقات ستاتا.

Windows Metafile (*.wmf)

Enhanced Metafile (*.emf)

Portable Network Graphics (*.png)

TIFF(*.tif)

PostScript (*.ps)

Encapsulated PostScript with or without TIFF preview (*.eps)

Portable Document File (*.pdf)

أنظمة التشغيل الأخرى مثل ماك Mac أو لينكس Linux توفر مجموعة مختلفة من الخيارات لحفظ الرسم البياني، بغض النظر عن أي تنسيق تريده لحفظ الملف. من الأفضل أن يتم حفظ نسخة من الرسم البياني بامتداد gph. حيث يمكن مع هذا الامتداد فتح الملف لاحقاً باستخدام برنامج ستاتا، ودمج الرسم وتغيير ألوانه وإعادة تنسيقه باستخدام الأمر **graph use** أو **graph combine** أو تحرير الرسم البياني باستخدام Graph Editor (الفصل الثالث من هذا الكتاب).

خلال كل مراحل التحليل في هذا الكتاب، ملف التسجيل *Monday1.smcl* سوف يقوم بحفظ النتائج، وأسهل طريقة لمعاينة ما تم حفظه في هذا الملف يتم بفتح الملف باستخدام نافذة المعاينة، ويكون ذلك باختيار:

File>Log > View > OK

ويمكننا طباعة محتويات ملف التسجيل بالضغط على أيقونة وهي في أعلى شريط المهام في نافذة المعاينة، وملفات التسجيل يتم إغلاقها تلقائياً عند إغلاق ستاتا، أو يتم ذلك قبل إغلاق ستاتا بالضغط على هذه الأيقونة

الموجودة في أعلى شريط المهام في نافذة ستاتا واختـر Close log file أو يمكنك كتابة الأمر **Log close** كما يمكن اختيار:

File>Log > Close

عند إغلاق ملف التسجيل *Monday1.smcl*، يمكن إعادة فتحـه ومـشاهدة محتوياته من خلال File> Log > View أو اضغط أيقونة بعد اسـتخدام ستاتا؛ ولإنشاء ملف مخرجات يمكن فتحه باستخدام أي معالج نـصوص أو ترجمته من ملف التسجيل الذي ينتهي بامتداد smcl (تنسيق سـتاتا لملفـات التسجيل) إلى log. (الصيغة العامة للملف النصي ASCII) قم بطباعـة الأمـر التالي :

.translate *Monday1.smcl Monday1.log*

أو يمكنك البدء بإنشاء ملف تسجيل بامتداد log. بدلاً من smcl. كما يمكنك أيضًا البدء والتوقف مؤقتاً في استعمال ملف التسجيل في أي لحظة باختيار:

File>Log > Suspend
File>Log > Resume

وأيقونة التسجيل الموجودة في شريط مهام ستاتا، يمكنها أيضًا القيام بكل هذه المهام.

المستندات والملفات المساعدة لبرنامج ستاتا :

Stata's Documentation and Help Files

دليل استخدام ستاتا 12 يتضمن 19 مجلدًا: دليل صغير الحجم عن بـدء استخدام ستاتا *Getting Started* (مثلاً: البدء باستخدام ستاتا مع ويندوز)، ودليل أكثر تركيزًا وهو دليل المستخدم *User's Guide* الذي يتكون من أربعة مجلدات شاملة تمثل مرجعاً شاملاً *Base Reference Manual*، ويحتوي على شرح عـن إدارة البيانات، والرسومات، والبيانـات الطوليـة، وبرمجـة المـصفوفات والاستدلال المتعدد والبرمجة وصياغة نماذج المعادلات الهيكلية، وبيانـات الاستبيانات، وجداول تحليل التعداد السكاني، وتحليل السلاسل الزمنية. كمـا أن دليل بدء استخدام ستاتا *Getting Started* يساعدك على تثبيت برنامج ستاتا، وإدارة ويندوز، وإدخال البيانات، والطباعة.. الخ، ولكن دليل المستخدم *User's*

Guide يحتوي على نقاش تفصيلي عن المواضيع العامة بما فيها مصادر البيانات، وحل المشاكل في ستاتا، ويجب ملاحظة أن دليل المستخدم يحتوي على الأوامر التي يجب أن يعرفها كل مستخدم، أما دليل المرجع *Base Reference Manual* يعرض قائمة أوامر ستاتا مرتبة هجائياً، وإدخالات الأوامر التي تتضمن اسم الأمر بالكامل، وشرح لكل المتغيرات والخيارات، وأمثلة وملاحظات تقنية عن الصيغ والشروط المنطقية، وقائمة مراجع للحصول على معلومات أكثر عن ستاتا؛ أما بخصوص إدارة البيانات والرسومات البيانية والبيانات الأخرى، فإنه تم تناولها في المراجع العامة. ولكن هذه المواضيع المعقدة تم تناولها بمزيد من التفاصيل والأمثلة في الكتب المتخصصة؛ وهناك أيضًا المرجع السريع والمؤشر *A Quick Reference and Index* الذي يحتوي على كل المواضيع التي تريد البحث عنها؛ بالرغم من أن الكتب التي تتحدث عن كيفية استخدام ستاتا موجودة بكثرة في أرفف المكتبات إلا أن نسخاً إلكترونية متكاملة على شكل ملفات PDF يمكن الوصول إليها ضمن برنامج ستاتا نفسه في أي وقت من خلال اختيار Help>PDF Documentation أو من خلال طباعة الأمر **help** ثم طباعة اسم الأمر.

عند القيام باستخدام ستاتا، فإنه من السهل الحصول على المساعدة وعرضها على الشاشة والتي بدورها يمكنها الاتصال مع أدلة استخدام ومراجع ستاتا، كما أن اختيار Help من شريط المهام يؤدي إلى ظهور قائمة منسدلة تحتوي على العديد من الخيارات، والتي من ضمنها أوامر Stata Command، ما هو الجديد ?What's new، تحديثات Online updates، مجلة ستاتا Stata Journal، البرامج المكتوبة بواسطة المستخدم user-written programs، ورابط للاتصال مباشرة بصفحة ستاتا على الإنترنت www.stata.com؛ اختيار Search يسمح لك كمستخدم بطباعة الكلمات التي تبحث عنها ضمن مستندات ستاتا Stata's documentation، كما يمكنك اختيار Contents (أو طباعة الأمر **help**) والذي يتيح لنا البحث والقيام ببعض المهام حسب تصنيفها، الأمر **help** يعتبر مفيدًا خصوصاً عند استخدام اسم الأمر بعده، فمثلاً طباعة الأمر **help correlate** يعرض شرحاً عن الأمر **correlate** في نافذة معاينة، وهذه المساعدة تعرض لك اسم الأمر وتركيبته وقائمة كاملة بالخيارات المتاحة مع الأمر، كما أنها أيضًا تتضمن

بعض الأمثلة ولكنها بدون شرح تفصيلي أو ملاحظات تقنية كما هو الأمر في أدلة استخدام ستاتا، والمساعدة التي يتم عرضها على الشاشة لها عدة مميزات عن تلك الموجودة بدليل استخدام ستاتا، حيث إن نافذة المعاينة تسمح لك بالبحث باستخدام الكلمات الرئيسة في الملفات الموجودة في صفحة ستاتا على الإنترنت، كما أن الروابط تنقلك مباشرة إلى الأوامر ذات العلاقة، إضافة إلى ذلك، فإن المساعدة على الشاشة قد تحتوي على بعض الملاحظات حول آخر التحديثات أو برامج ستاتا غير الرسمية، والتي يمكنك تحميلها من صفحة ستاتا أو من المستخدمين الآخرين لبرنامج ستاتا.

البحث عن المعلومات* : Searching for Information

عند اختيار Help > Search > Search documentation and FAQs يزودك بطريقة واضحة للبحث عن معلومات معينة في أدلة استخدام ستاتا أو في صفحة الأسئلة المتكررة، والصفحات الأخرى، كما يمكنك أيضاً البحث مباشرة في مجلة ستاتا *Stata Journal*، ونتائج البحث تظهر في نافذة تحتوي على روابط عند النقر على أي منها سوف ينقلك إلى معلومات أكثر.

ويقوم الأمر **search** بنفس المهام، وأحد الاستخدامات الخاصة للأمر **search** هو البحث عن معلومات أكثر تفصيلاً عن أمر معين لا يعمل كما يفترض، ولكن بدلاً من ذلك يمكنك القيام بالنقر على الرقم الذي يظهر في رسالة الخطأ للحصول على هذه المعلومات. فعلى سبيل المثال **table** هو أحد أوامر ستاتا، ولكنه يتطلب معلومات محددة عن ماذا نريد في الجدول، فإذا قمنا بطباعة أمر **table** بذاته بدون أي شيء آخر، فإن ستاتا يعرض الرسالة الخطأ التالية: "return code" r(100)

```
.table
varlist required
r(100);
```

* كل معلومات الدعم الفني والمساعدة في برنامج ستاتا تُعرض باللغة الإنجليزية.

عند النقر على ;r(100) فإنه سوف يتم نقلك إلى معلومات مفيدة، والتي يمكننا كذلك الحصول عليها بواسطة كتابة الأمر **search rc 100** أو طباعة **help search** للحصول على معلومات حول أي أمر لستاتا.

شركة ستاتا : Stata Corp

العنوان البريدي لشركة ستاتا هو:

Stata Corp
4905 Lakeway Drive
College Station, TX 77845 USA

وأرقام هواتف الشركة هي:

من الولايات المتحدة	001-782-8272 أو (1-800-STATAPC)
من كندا	001-782-8272
باقي دول العالم	001-979-696-4600
فاكس	001-979-696-4601

لشراء أو الحصول على رخصة استعمال أو معلومات التحديث، يمكنك الاتصال بشركة ستاتا عن طريق البريد الإلكتروني:

service@stata.com

أو زيارة صفحة ستاتا على الإنترنت

www.stata.com

كما أن صفحة ستاتا على الإنترنت، بها جزء خاص بالإعلام والإصدارات الصحفية، بما في ذلك المنشورات والبيانات التي تم استخدامها في أمثلة بعض الكتب.

www.stata-press.com

وهناك مجلة ستاتا المحكّمة، والتي أصبحت مصدرًا مهماً وموقعها www.stata-journal.com

صفحة ستاتا الرئيسة www.stata.com تزود المستخدم بمعلومات مكثفة، حيث إنها تبدأ بصفحات تشرح بالتفصيل منتجات ستاتا، وكيفية شراء منتجات ستاتا، والعديد من طرق دعم العملاء مثل:

FAQs وهي أسئلة متكررة مع أجوبتها عن بعض المشكلات التي لا توجد في أدلة استخدام ستاتا، الأسئلة تتنوع من أسئلة مثل "كيف تستطيع تحويل ملفات التطبيقات الإحصائية الأخرى إلى تنسيق ملفات ستاتا؟ إلى أسئلة ذات طبيعة تقنية أكثر مثل "كيف يمكنك وضع بعض المحددات على قيمة صفر باستخدام الأمر heckman مع ml؟

التحديثات Updates : أغلب إصدارات ستاتا المرخّصة يمكن تحديثها مجاناً للمستخدمين، وهذا يوفر طريقة سريعة وبسيطة للحصول على أحدث التطورات والإصلاحات.. الخ للإصدار الذي تقوم باستخدامه، فبدلاً من الذهاب إلى صفحة ستاتا على الإنترنت، يمكنك التأكد ما إذا كانت هناك أي تحديثات جديدة لبرنامج ستاتا الذي تقوم باستخدامه عن طريق طباعة الأمر

`.update query`

الدعم الفني: بالإمكان الحصول على الدعم الفني بواسطة إرسال بريد إلكتروني إلى tech-support@stata.com، والردود في العادة تكون مفيدة، ولكن قبل إرسال بريد الكتروني يُفترض أن تقوم بمحاولة البحث عن حل لأي مشكلة من خلال FAQs الأسئلة المتكررة.

التدريب: بالإمكان التسجيل في بعض الدورات التدريبية على الإنترنت لبعض المواضيع المتخصصة في ستاتا مثل مقدمة في ستاتا، مقدمة إلى برمجة ستاتا، أو برمجة ستاتا المتقدمة.

أخبار ستاتا: تحتوي على معلومات حول مميزات البرنامج والدورات التدريبية وآخر القضايا الخاصة بمجلة ستاتا ومواضيع أخرى.

المنشورات: هناك روابط عن مجلة ستاتا، ومستندات وأدلة استخدام ستاتا، وكذلك الكتب المعروضة للبيع عن ستاتا، وأحدث مراجع ستاتا، بالإضافة إلى برامج الدعم لمؤلفي كتب جديدة عن ستاتا؛ كما أن صفحة ستاتا

تستضيف منتدى خاصاً بالنقاش حول مجلة ستاتا وكتب ستاتا، وروابطها وهذه يمكن الاطلاع عليها من خلال الرابط التالي:

blgo.stata.com

مستخدمو مواقع التواصل الاجتماعي ربما يجدون أنه من الممتع والمفيد متابعة ستاتا على تويتر www.twitter.com أو اضغط معجب لصفحة ستاتا على الفيس بوك www.facebook.com

مجلة ستاتا : The Stata Journal

منذ سنة 1991 وحتى سنة 2001، كان هناك إصدار نصف شهري يسمى التقرير التقني لستاتا *Stata Technical Bulletin* وتم استخدامه كوسيلة لنشر الأوامر الجديدة وتحديثات برنامج ستاتا للمستخدمين، وفي نهاية كل سنة يتم تجميع هذه التقارير ونشرها في كتاب يسمى إعادة طباعة التقارير التقنية لستاتا *Stata Technical Bulletin Reprints* ويمكن شراء هذا الكتاب من شركة ستاتا مباشرة؛ ومع تطور الإنترنت أصبح الاتصال بالمستخدمين أمرًا سهلاً، وملفات البرامج يمكن بسهولة تحميلها من الإنترنت من مصادر بعيدة، وأصبح التقرير النصف الشهري المطبوع لا يفي بحاجات المستخدمين ومعدي البرامج، لذلك فقد تم تغيير التقرير التقني إلى شكل آخر أحدث.

مجلة ستاتا *The Stata Journal* والتي تم إصدارها لتفي بالحاجات المتزايدة للعدد المتزايد لمستخدمي ستاتا؛ مجلة ستاتا تشبه التقرير التقني من حيث إنها تحتوي على مقالات تشرح أوامرستاتا الجديدة، والإصدار غير الرسمي لبعض الأوامر، والتي تم إعدادها بواسطة العاملين في شركة ستاتا؛ التركيز الأساسي للمجلة ليس على الأوامر الجديدة فقط، وإنما أيضاً على مقالات محكّمة عن الإحصاء، ومراجعات للكتب، وملاحظات عن استخدام ستاتا، وعدد من الأعمدة المفيدة والتي منها عمود "ستاتا تتحدث" *Speaking Stata* لكاتبه نيكولاس كوكس Nicholas J. Cox حول الاستخدام الفعال للغة برامج ستاتا، فمجلة ستاتا هي للمستخدمين المبتدئين والمحترفين على حد سواء، فمثلاً إليك محتويات مجلة ستاتا في عددها الصادر في يونيو 2012:

المقالات والأعمدة

"A robust instrumental-variables estimator," R. Desbordes, V. Verardi

"What do hypotheses do "nonparametric" two-group tests actually test?" R.M. Conroy

"From resultsses to resultstables in Stata," R.B. Newson

"Menu-driven X-12-ARIMA seasonal adjustment in Stata," Q. Wang, N. Wu

"Faster estimation of a discrete-time proportional hazards model with gamma frailty," M.G. Farnworth

"Threshold regression for time-to-event analysis: The stthreg package," T. Xiao, G.A. Whitmore, X. He, M.-L.T. Lee

"Fitting nonparametric mixed logit models via expectation-maximization algorithm," D. Pacifico

"The S-estimator of multivariate location and scatter in Stata," V. Verardi, A. McCathie

"Using the margins command to estimate and interpret adjusted predications and marginal effects," R. Williams

"Speaking Stata: Trnsforming the time axis," N.J. Cox

ملاحظات وتعليقات

"Stata tip 108: On adding and constraining," M.L. Buis

"Stata tip 109: How to combined variables with missing values," P.A. Lachenbruch

"Stata tip 110: How to get the optimal k-means cluster solution," A. Makles

تحديثات البرنامج

يتم إصدار مجلة ستاتا بشكل ربع سنوي، والاشتراكات في هذه المجلة يمكن أن تتم بزيارة رابط المجلة www.stata-journal.com، كما يحتوي الموقع على قائمة أرشيف تتضمن كل الأعداد السابقة من المجلة، والتي يمكنك شراءها كلاً على حدة، كما يمكنك تحميل مقالات السنوات الثلاث الماضية مجاناً، وتم إصدار عدد خاص من المجلة بمناسبة الذكرى العشرين (العدد 5 المجلد 1 سنة 2005) احتوى العدد على مقالات حول تطوّر ستاتا، وكتاب حول ستاتا بعنوان "تاريخ قصير للإحصاء مع ستاتا" *A short history of Statistics with Stata*

كتب عن استخدام ستاتا : Books Using Stata

بالإضافة إلى أدلة استخدام ستاتا التي تأتي مع البرنامج نفسه، هناك نمو كبير في عدد الكتب التي تشرح ستاتا واستخداماته وتقنيات التحليل باستخدام ستاتا. هذه الكتب تحتوي على مقدمة عامة، والتطبيقات المتعلقة به مثل العلوم الاجتماعية أوالاقتصاد القياسي، كما يركز على النصوص المتعلقة بتحليل الاستبيانات، وبيانات التجارب المعملية، والمتغيرات المستقلة المصنفة ومواضيع أخرى.

كما أن مكتبة بيع الكتب على صفحة ستاتا لديها قائمة حديثة مع شرح لمحتويات الكتب على الرابط التالي :

http://www.stata.com/bookstore/

هذه المكتبة تزودك بمكان رئيس للتعرّف على أحدث الكتب المتعلقة ببرنامج ستاتا من مختلف ناشري هذه الكتب. فعلى سبيل المثال، هناك العديد من الكتب منها:

A Gentle Introduction to Stata, A.C. Acock

Using Stata for Principles of Econometrics, L.C. Adkins, R.C. Hill

An Introduction to Modern Econometrics Using Stata, C.F. Baum

Applied Microeconometrics Using Stata, A.C. Cameron, P.K. Trivedi

Event History Analysis with Stata, H-P. Blossfeld, K. Golsch, G.Rohwer

An Introduction to Survival Analysis Using Stata, M. Cleves, W. Gould, R. Gutierrez, Y. Marchenko

Statistical modeling for Biomedical Researchers, W.D. Dupont

Maximum Likelihood Estimation with Stata, W. Gould, J. Pitblado, B. Poi

Statistics with Stata, L.C. Hamilton

Generalized Linear Models and Extensions, J.W. Hardin, J.M. Hilbe

Negative Binomial Regression, J.M. Hible

A Short Introduction to Stata for Biostatistics, M. Hills, B.L. De Stavola

Applied Survival Analysis: Regression Modeling of Time to Event Data, D.W. Hosmer, S. Lemeshow, S. May

Applied Econometrics for Health Economists, A. Jones

Applied Health Economics, A. Jones, N. Rice, T.B. d'Uva, S. Balia

An Introduction to Stata for Health Researchers, S. Juul, M. Frydenberg

Data Analysis Using Sata, U. Kohler, F. Kreuter

Sampling of Populations: Methods and Applications, P.S. Levy, S. Lemeshow

Working in Data Analysis Using Stata, J.S. Long

Regression Models for Categorical Dependent Variables Using Stata, J.S. Long, J. Freese

A visual Guide to Stata Graphics, M. Mitchell

Data Management Using Stata: A Practical Handbook, M. Mitchell

Interpreting and Visualizing Regression Models Using Stata,M. Mitchell

Seventy-six Stata Tips, H.J. Newton, N. J. Cox editors

Analyzing Health Equity Using Household Survey Data, O. O'Donnell and others

A Stata Companion to Political Analysis, P.H. Pollock III

A Handbook of Statistical Analyses Using Stata, S. Rabe-Hesketh, B. Everitt

Multilevel and Longitudinal Modeling Using Sata, S. Rabe-Hesketh, A. Skrondal

Managing Your Patients? Data in the Neonatal and Pediatric ICU, J. Schulman

Epidemiology: Study Design and Data Analysis, M. Woodward



الفصل الثاني

إدارة البيانات

Data Management

أول خطوة في تحليل البيانات تتضمن تنظيم صفوف البيانات، في شكل يمكن به معرفة مفهوم برنامج ستاتا. يمكننا إدخال بيانات جديدة إلى ستاتا باستخدام عدة طرق: طباعة البيانات باستخدام لوحة المفاتيح، استيراد البيانات من برامج أخرى مثل برنامج مايكروسوفت إكسيل، قراءة البيانات من ملف نصي أو ملف ASCII يحتوي على صفوف البيانات، لصق البيانات من جداول في المحرر، أو استخدام برنامج يقوم بتحويل البيانات، ترجمة البيانات مباشرة من ملف نظام تم إنشاؤه بواسطة برنامج جداول إليكترونية آخر، أو قاعدة بيانات، أو برنامج إحصائي آخر. عند إدخال البيانات في برنامج ستاتا يمكننا حفظها في ملف بتنسيق ستاتا، بحيث يسهل إعادة فتحه وتحديثه مستقبلاً باستخدام ستاتا.

إدارة البيانات تتضمن: المهام الأساسية التي عن طريقها يتم إنشاء ملف البيانات وتحرير هذه البيانات وتصحيح الأخطاء، وتحديد البيانات المفقودة، وإضافة مستندات داخلية مثل المتغيرات والقيم والأسماء. كما أن إدارة البيانات تتضمن: العديد من الوظائف التي تتطلبها عملية التحليل مثل إضافة مشاهدة جديدة، أو متغير جديد، وإعادة التنظيم، واختيار عينة من البيانات، وفصل وتجميع أو إلغاء بيانات، وتحوير أنواع المتغيرات، وإنشاء متغيرات من خلال شروط منطقية أو رياضية معينة. وعندما تكون عملية إدارة البيانات متكررة أو معقدة، فإن مستخدمي ستاتا يمكنهم كتابة برنامج خاص بهم لإجراء العمليات بشكل تلقائي، حيث إن ستاتا معروف بقدرته التحليلية الفائقة، فيمكنه التعامل مع بيانات متنوعة، وكذلك إدارة مميزاتها.

دليل المستخدم لبرنامج ستاتا *User's Guide* يعطي نظـرة عامـة علـى الطرق المختلفة لإدخال البيانات، ثم بعد ذلك يشرح تـسع قواعـد لتحديـد الطريقة المناسبة. إدخال البيانات وتحريرها والعديد من المهـام الأخـرى، سوف تتم مناقشتها في هذا الفصل، ويمكن القيام بهذه المهام من خلال قائمة Data، وهي تحتوي على القوائم الفرعية التالية:

- Describe data
- Data Editor
- Create or change data
- Variables Manager
- Data utilities
- Sort
- Combine datasets
- Matrices, Mata language
- Matrices, ado language
- Other utilities

أمثلة عن الأوامر : Example Commands

```
.append using olddata
```

قم بقـراءة ملف البيانات المخزن مسبقاً *olddata.dta* ثم قـم بإضـافة مفـرداته إلى ملف البيانات المستخـدم حالياً، بعد ذلك قـم بكتابـة الأمـر **save *newdata*, replace** فسوف يتم حفظ البيانات الموحدة في ملف جديـد اسمه *newdata.dta*

```
.browse
```

هذا الأمر يقوم بفتح جدول إلكتروني يشبه متصفح البيانـات لمـشاهدة بيانات، والمتصفح يشبه محرر البيانات، ولكن ليس لديه القدرة على تحريـر البيانات، فلايوجد هناك خطر على تغيير البيانات بـشكل غيـر مقـصود؛ ويمكنك القيام بنفس الأمر بالنقر على أيقونة

```
.browse year month extent if year> 1999
```

هذا الأمر يفتح متصفح البيانات، ويعرض فقط المتغيرات *year, month, extent*، كما يعرض المشاهدات التي تم إدخالها في سنة *year* 1999 وما بعدها، في هذا المثال **if** تمثل الشرط المنطقي والذي يجب استخدامه لجعل العديد من أوامر ستاتا أكثر تركيزاً.

```
.compress
```

يقوم هذا الأمر بشكل تلقائي بضغط البيانات، وتحويل كل المتغيرات إلى أكثر أشكال التخزين كفاءة للمحافظة على السعة التخزينية للجهاز، ثم بعد ذلك قم بكتابة الأمر **save *filename*, replace** وهذا سوف يحفظ التغييرات بشكل دائم.

```
.drawnorm z1 z2 z3, n(5000)
```

يقوم هذا الأمر بإنشاء بيانات تحتوي على 5000 مشاهدة و3 متغيرات عشوائية هي *z1, z2, z3*. وهذه البيانات غير مرتبطة، وتتوزع توزيعاً طبيعياً؛ وهناك خيارات يمكن استخدامها لتحديد المتوسط، والانحراف المعياري ومصفوفات الارتباط.

```
.dropmiss
```

وهذا الأمر يقوم تلقائياً بحذف أي متغير يحتوي على قيم مفقودة لكل المشاهدات، هذا الأمر يعتبر مفيدًا عند العمل مع كمية كبيرة من البيانات، حيث إن بعض المتغيرات الأصلية غير متعلقة بأي مشاهدة من المشاهدات المتبقية؛ وكتابة الأمر **dropmiss, obs** سوف يقوم باستبعاد أي مشاهدات لها قيم مفقودة؛ يجب ملاحظة أن **dropmiss** أمر تم برمجته بواسطة أحد المستخدمين، فهو لايأتي ضمن الأوامر الافتراضية في ستاتا، لذا يجب تحميله من الإنترنت وتنصيبه في ستاتا، ويمكنك الحصول على رابط التحميل عن طريقة طباعة الأمر **findit dropmiss**.

```
.edit
```

يقوم بفتح نافذة بها جدول البيانات، حيث يمكنك إدخال وتحرير البيانات، ويمكن القيام بنفس الأمر بالنقر على أيقونة تحرير من شريط المهام.

```
.edit year month extent
```

يقوم بفتح نافذة تحرير البيانات للمتغيرات *year, month, extent* على التوالي، ويمكنك تعديل البيانات حسب الحاجة.

```
.encode stringvar, gen (numvar)
```

يقوم هذا الأمر بإنشاء متغير جديد اسمه *numvar* مع توصيفه بقيم رقمية بناءً على متغير نصي (غير رقمي) اسمه *stringvar*

```
.format rainfall %8.2f
```

يقوم هذا الأمر بإنشاء تنسيق عرض ثابت (**f**) للمتغير الرقمي *rainfall* بحيث يكون عرض عمود المتغير 8 أعمدة، ودائماً يظهر رقمان بعد الفاصلة العشرية؛ هذا الأمر يؤثر فقط في كيفية عرض البيانات، ولا يؤثر في البيانات نفسها.

```
.generate newvar = (x+y)/100
```

يقوم الأمر بإنشاء متغير جديد اسمه *newvar* وهو يساوي x مضافاً إليها y والناتج مقسم على 100.

```
.generate newvar = runiform( )
```

يقوم هذا الأمر بإنشاء متغير جديد من قيم تم أخذها من توزيع عشوائي منتظم، تتراوح مابين 0 وتقريباً 1 وتُكتب [0,1). وللحصول على معلومات عن الدوال التي تقوم بإنشاء بيانات عشوائية من التوزيع الطبيعي، والتوزيع الثنائي، وتوزيع χ^2، وتوزيع جاما، وتوزيع بواسون، والتوزيعات الأخرى قم بكتابة الأمر **help random**.

```
.import excel filename.xlsx, sheet ("mean")
 cellrange (a15:n78) firstrow
```

يقوم هذا الأمر باستيراد ورقة إكسيل داخل برنامج ستاتا، والخيار الثاني في هذا المثال، يوضح أن اسم الورقة هو "mean"، والتي تحتوي على بيانات في الخلايا من A15 إلى N78. والصف الأول من هذه الخلايا يتضمن أسماء المتغيرات.

```
.infile x y z using data.raw
```

هذا الأمر يقوم بقراءة الملف المسمى *data.raw* والذي يتضمن بيانات ثلاثة متغيرات هي *x, y, z* والقيم المتعلقة بهذه المتغيرات الثلاثة تم الفصل بينها بمسافة واحدة (يمكن استخدام الفراغات ومسافة TAB والفواصل والخطوط)، المسافات تحدد القيم المفقودة للمتغيرات الرقمية، والتي يجب أن يتم تمثيلها بواسطة الفترات وليس الفراغات؛ استخدام الفاصلة كحد فاصل بين القيم والفراغات أو فاصلتين متتاليتين للقيم المفقودة. وللحصول على تفاصيل أكثر عن الأوامر التي تقوم بقراءة البيانات من ملفات مختلفة واستيرادها داخل ستاتا قم بطباعة الأمر **help infiling**.

```
.list
```

يقوم بعمل قوائم للبيانات في التنسيق الافتراضي للجداول، عند العمل مع حجم بيانات ضخم، فإن تنسيق الجداول يصبح مهمة صعبة، لذا فإن استخدام الأمر **list** والأمر **display** يعطي نتائج مفيدة. للحصول على معلومات عن الخيارات الأخرى المتوافرة مع هذا الأمر، قم طباعة الأمر **help list**؛ كما أن محرر البيانات Data Editor ومتصفح البيانات Data Browser يوفر العديد من الخيارات للمعاينة حسب الغرض المطلوب.

```
.list x y z in 5/20
```

يقوم هذا الأمر بعمل قائمة بالمتغيرات *x, y, z*، والقائمة تتضمن المشاهدات من المشاهدة رقم 5 إلى المشاهدة رقم 20 حسب تسلسل إدخالها.

```
.merge 1:1 id using olddata
```

يقوم هذا الأمر بقراءة البيانات المخزنة مسبقاً بملف *olddata.dta* للمتغير *id* ثم يقارن المشاهدات الموجودة بالملف مع تلك المشاهدات الموجودة بالملف الحالي المفتوح، حيث تتم مقارنة المشاهدات واحدة بواحدة، ويجب أن يتم ترتيب بيانات الملفين وفقاً للمتغير *id*.

```
.mvdecode var3-var62, mv(97=.\98=.a\99=.b)
```

بالنسبة للمتغيرات من var3 وحتى var62 قم باعتبار القيم 97، 98، 99 على أنها قيم مفقودة. في هذا المثال، نحن نستخدم ثلاثة رموز مختلفة لتعريف القيم المفقودة وهي .a, .b ,. هذه القيم يمكن الاستعاضة عنها باستخدام

أسباب نقص هذه القيم مثل كتابة غير قابلة للتطبيق Not applicable، غير معروف Don't know، رفض الإجابة Refused to answer؛ إذا كان المطلوب هو استخدام رمز واحد للقيم الناقصة. إذن يمكننا تحديد رمز واحد للقيم المفقودة، فإذا قمنا باستخدام (.) فإن الأمر يمكن كتابته (.=99 98 97)mv

```
.replace oldvar = 100 * oldvar
```

يقوم الأمر باستبدال قيم المتغير *oldvar* بقيم جديدة تم حسابها بضرب القيم القديمة للمتغير القديم في 100.

```
.sample 10
```

يقوم الأمر بحذف كل المفردات في الملف الحالي المفتوح باستثناء 10% سوف يتم الإبقاء عليها كعينة عشوائية؛ فبدلاً من اختيار نسبة محددة يمكننا اختيار عدد محدد من الحالات، فمثلاً **sample 55, count** سوف يقوم بحذف كل المشاهدات باستثناء عينة حجمها 55 مشاهدة.

```
.save newfile
```

يتم حفظ البيانات في الملف المفتوح حالياً والمسمى *newfile.dta*. إذا كان الملف *newfile.dta* موجودًا مسبقاً، وتريد الكتابة على الإصدار السابق، يمكنك كتابة الأمر ***save newfile*, replace**، أو يمكنك القيام بنفس المهمة عن طريق قائمة File، ولحفظ الملف *newfile.dta* بتنسيق ستاتا 9 قم بكتابة الأمر **saveold *newfile*** أو قم باختيار File > Save As > Save as type.

```
.sort x
```

يقوم هذا الأمر بترتيب مشاهدات المتغير *x* تصاعدياً من أقل قيمة إلى أعلى قيمة، وتظهر القيم المفقودة في نهاية القائمة، لأن برنامج ستاتا يعتبر القيم المفقودة أعلى قيم. للحصول على معلومات أكثر حول أمر الترتيب العام، قم بطباعة الأمر **help gsort** حيث يوفر معلومات عن خيارات الترتيب تنازلياً، أو تصاعدياً، أو وضع القيم المفقودة في البداية.

```
.tabulate x if y> 65
```

يقوم هذا الأمر بإنشاء جدول تكراري للمتغير *x* فقط باستخدام المشاهدات التي يناظرها المتغير *y* يزيد على 65؛ استخدام **if** يشبه تماماً استخدامها مع معظم أوامر ستاتا الأخرى.

```
.use oldfile
```

يقوم هذا الأمر باستخدام ملف البيانات المخزن مسبقاً باسم *oldfile.dta*. إذا كان هناك ملف بيانات مفتوح حالياً، وتريد إلغاء هذه البيانات بدون حفظها فيمكنك طباعة الأمر **use *oldfile*, clear** أو يمكنك القيام بنفس المهمة عن طريق اختيار File > Open أو الضغط على أيقونة

إنشاء بيانات بطباعتها في نافذة Data :

Creating a New Dataset by Typing in Data

البيانات التي تم حفظها مسبقاً في ملف ستاتا، يمكن استخدامها بفتح الملف المخزن بأمر **use *filename*** أو من خلال القوائم. في هذا الجزء، سوف يتم شرح المهارات الأساسية لإنشاء ملف بيانات باستخدام برنامج ستاتا، حيث يمكننا ببساطة طباعة البيانات يدوياً في نافذة تحرير البيانات Data Editor، أو إدخال البيانات يدوياً. وهذا يعتبر عملياً عندما يكون حجم البيانات بسيطاً أو البيانات ليست مخزنة على وسيلة إلكترونية مثل ملف إكسيل، ولكن هناك العديد من الطرق الأخرى لإدخال البيانات.

الجدول (1.2) يعرض بعض المعلومات عن الولايات والأقاليم الكندية، وهذه البيانات يمكن استخدامها لتوضيح كيفية إدخال البيانات يدوياً. تم الحصول على هذه البيانات من لجنة الإشراف الإقليمية البلدية الاتحادية على صحة السكان لسنة 1996، أحدث أقاليم كندا (Nunavut) لم يتم إدراجها ضمن البيانات، وذلك لأنها كانت جزءًا Wh من الأقاليم الشمالية الغربية حتى سنة 1999.

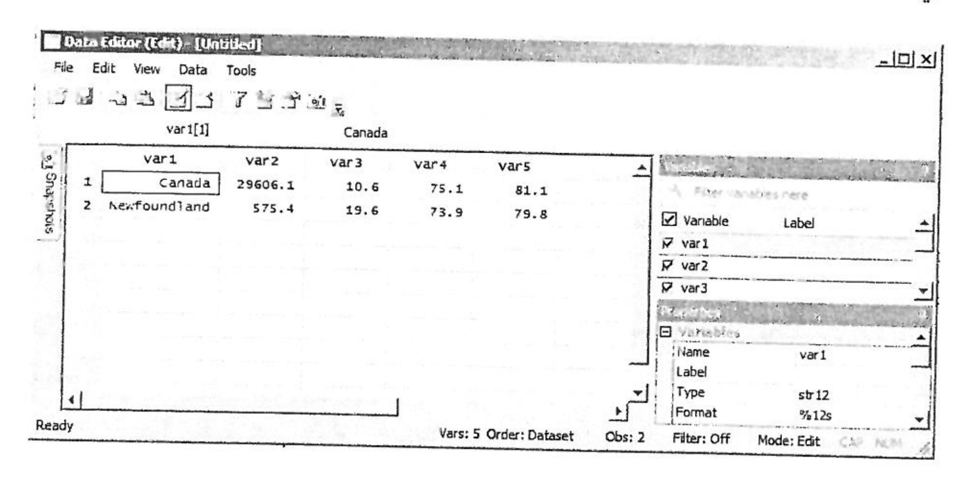
جدول (1.2) : بيانات عن كندا وولاياتها

المكان Place	عدد السكان 1995 (بالآلاف) 1995 Pop. (1000's)	معدل البطالة (نسبة) Unemployment Rate (percent)	العمر المتوقع للذكور (سنوات) Male Life Expectancy (years)	العمر المتوقع للإناث (سنوات) Female Life Expectancy (years)
Canada	29606.1	10.6	75.1	81.1
Newfoundland	575.4	19.6	73.9	79.8
Prince Edward Island	136.1	19.1	74.8	81.3
Nova Scotia	937.8	13.9	74.2	80.4
New Brunswick	760.1	13.8	74.8	80.6
Quebec	7334.2	13.2	74.5	81.2
Ontario	11100.3	9.3	75.5	81.1
Manitoba	1137.5	8.5	75.0	80.8
Saskatchewan	1015.6	7.0	75.2	81.8
Alberta	2747.0	8.4	75.5	81.4
British Columbia	3766.0	9.8	75.8	81.4
Yukon	30.1	-	71.3	80.4
Northwest Territories	65.8	-	70.2	78.0

أسهل طريقة لإنشاء ملف بيانات من بيانات مطبوعة على ورق مثل الجدول (1.2) هي استخدام نافذة تحرير البيانات Data Editor والتي تظهر من خلال النقر على أو من خلال اختيار Window > Data Editor من قائمة ستاتا أو من خلال طباعة الأمر **edit**، ثم بعد ذلك قم بطباعة قيم كل متغير، الأعمدة تتم تسميتها *var1*، *var2*، الخ؛ لذلك فإن *var1* يحتوي على بيانات المكان place، و*var2* يحتوي على بيانات عدد السكان وهكذا.

يمكننا وضع أسماء أكثر تعبيراً عن المتغيرات وذلك بالنقر على أعلى كل عمود (مثلاً المتغير *var1*) في نافذة تحرير البيانات، تم كتابة الاسم الجديد في مربع حوار النتائج، بالرغم من أن اسم المتغير يمكن أن يحتوي على 32 حرفاً، إلا أنه من المفضل أن يتكون من 8 حروف أو أقل، كما يمكننا أيضاً أن نضع وصفاً يتضمن شرحاً مختصرًا لاسم المتغير، فمثلاً

var2 (عدد السكان population) يمكن تسميته *pop* ويعطى شرحاً مختصرًا في حقل variable lable كالآتي "Population in 1000s, 1995".



إعادة تسمية المتغيرات يمكن أن تتم في نافذة تحرير البيانات Data Editor، أو من خلال استخدام الأوامر **rename** و**labelvariable** كما يلي:

```
.rename var2pop
.label variable pop "Population in 1000s, 1995"
```

الفضاءات التي تم تركها خالية، مثل معدلات البطالة unemployment rates للولايات Yukon وNorthwest Territories سوف يعتبرها ستاتا تلقائياً قيماً مفقودة، ويمكننا إغلاق نافذة تحرير البيانات في أي وقت، وحفظ البيانات في الجهاز؛ عند الضغط على أيقونة تحرير البيانات أو اختيار Data > Data Editor أو طباعة الأمر **edit** سوف يتم إرجاعك إلى نافذة تحرير البيانات.

إذا كانت القيمة الأولى التي تم إدخالها لمتغير ما رقم – مثل عدد السكان population ومعدل البطالة unemployment rate والعمر المتوقع life expectancy – فإن ستاتا سوف يعتبر العمود متغيرًا رقمياً، ولاحقاً سوف يسمح فقط بقيم رقمية لذلك المتغير. القيم الرقمية يمكن أن تبدأ بعلامة زائد أو ناقص، ويمكن أن تتضمن علامة عشرية أو تعبيرات رياضية. فمثلاً يمكننا التعبير عن عدد سكان كندا كرقم 2.96061e+7 وهذا يساوي 2.96061

10^7 × أي حوالي 29.6 مليون نسمة؛ يُفترض ألا تحتوي الأرقام علـــى أي فواصل مثل 29,606,100 (ويمكن استخدام الفواصل كعلامات عشرية). وإذا حدث هذا، فإن ستاتا سوف يعتبر المتغير كمتغير نـــصـــي ولـــيس متغيـــراً رقمياً.

إذا كانت القيمة الأولى التي تم إدخالها لمتغير ما حـــرفاً وليســـت رقماً – مثل اسم المكان في الجدول السابق – فإن ستاتا سوف يعتبر ذلك المتغيـــر متغيرًا نصياً. قيم المتغير النصي يمكن أن تتضمن حروفاً وأرقاماً وتعبيرات ومسافات حتى 244 حرفاً. المتغيرات النـــصية يمكـــن أن تحفـــظ أســـماء وعلامات تنصيص وأي معلومات توضيحية أخـــرى، كمـــا يمكـــن وضـــع المتغيرات النصية في جداول ويمكن عدّها، ولكن لا يمكن تحليلها باســـتخدام المتوسط الحسابي والارتباط أو أغلب العمليات الإحصائية الأخـــرى، وفـــي نافذة محرر البيانات Data Editor أو نافذة متصفح البيانـــات Data Browser تظهر المتغيرات النصية بلون أحمر يميّزها عن المتغيرات الرقميـــة التـــي تظهر بلون أسود أو أسماء المتغيرات الرقمية (الزرقاء).

بعد طباعة المعلومات من الجدول (1.2) سوف نقوم بإغلاق نافذة محرر البيانات Data Editor وحفظ البيانات باسم *Canada1.data* وذلك بالأمر

```
.save Canada1
```

برنامج ستاتا سوف يقوم تلقائياً بإضافة امتداد dta. إلى ملف البيانات ما لم تقم بكتابة أي شيء آخر، إذا كنا قد قمنا بتخزين الملف مسبقاً بنفس الاسم، فإنه من المحتمل أن يتم تخزين الإصدار الأحدث من نفس الملـــف، وذلـــك بطباعة الأمر التالي:

```
.save, replace
```

عند هذه النقطة ملف البيانات سوف يظهر كما يلي:

```
.describe
```

```
Contains data from C:\data\Canada1.dta
  obs:            13                          Canadian dataset 1
 vars:             5                          4 Jul 2012 11:21
 size:           481

              storage  display     value
variable name   type   format      label      variable label

place           str21  %21s                   Place name
pop             float  %9.0g                  Population in 1000s, 1995
unemp           float  %9.0g                  % 15+ population unemployed, 1995
mlife           float  %9.0g                  Male life expectancy years
flife           float  %9.0g                  Female life expectancy years

Sorted by:
```

.list

	place	pop	unemp	mlife	flife
1.	Canada	29606.1	10.6	75.1	81.1
2.	Newfoundland	575.4	19.6	73.9	79.8
3.	Prince Edward Island	136.1	19.1	74.8	81.3
4.	Nova Scotia	937.8	13.9	74.2	80.4
5.	New Brunswick	760.1	13.8	74.8	80.6
6.	Quebec	7334.2	13.2	74.5	81.2
7.	Ontario	11100.3	9.3	75.5	81.1
8.	Manitoba	1137.5	8.5	75	80.8
9.	Saskatchewan	1015.6	7	75.2	81.8
10.	Alberta	2747	8.4	75.5	81.4
11.	British Columbia	3766	9.8	75.8	81.4
12.	Yukon	30.1	.	71.3	80.4
13.	Northwest Territories	65.8	.	70.2	78

.summarize

Variable	Obs	Mean	Std. Dev.	Min	Max
var1	0				
pop	13	4554.769	8214.304	30.1	29606.1
var3	11	12.10909	4.250048	7	19.6
var4	13	74.29231	1.673052	70.2	75.8
var5	13	80.71539	.9754027	78	81.8

اختبار مثل هذه المخرجات يعطينا الفرصة لرؤية الأخطاء التي يُفترض أن يتم تصحيحها، فمثلاً جدول **summarize** يوضح مجموعة من الأرقام المفيدة للمراجعة، وهذه الأرقام تتضمن عدد المشاهدات الرقمية (دائماً 0 للمتغيرات النصية) وأقل وأعلى قيمة لكل متغير، التفسير الموضوعي للإحصائيات المختصرة عند هذه النقطة سابق لأوانه، لأن البيانات تحتوي على مشاهدة واحدة (كندا) والتي تمثل دمجاً لباقي 12 ولاية.

الخطوة التالية هي إعطاء البيانات توثيقاً ذاتياً أكبر من خلال إعطاء المتغيرات أسماء أكثر وضوحاً، وذلك كما يلي:

```
.rename var1 place
.rename var3 unemp
.rename var4 mlife
.rename var5 flife
```

ويمكن القيام بنفس المهمة في خطوة واحدة كما يلي:

```
.rename (var1 var2 var4 var5)(place unemp mlife
 flife)
```

يتيح لنا برنامج ستاتا إضافة عدة أنواع من التوصيفات إلى البيانات، الأمر **label data** يشرح البيانات كذلك، بينما الأمر **label variable** يشرح كل متغير على حدة، وذلك كما يلي:

```
.label data "Canadian dataset 1"
.label variable place "Place name"
.label variable unemp "%15+ population
 unemployed, 1995"
.label variable mlife "Male life expectancy
 years"
.label variable flife "Female life expectancy
 years"
```

بإعطاء وصف للبيانات والمتغيرات، سوف نحصل على بيانات توضح نفسها بنفسها من خلال التوصيفات:

```
. describe

Contains data from C:\data\Canada1.dta
  obs:           13                          Canadian dataset 1
 vars:            5                          4 Jul 2012 11:21
 size:          481
--------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
--------------------------------------------------------------------------------
place           str21  %21s                   Place name
pop             float  %9.0g                  Population in 1000s, 1995
unemp           float  %9.0g                  % 15+ population unemployed, 1995
mlife           float  %9.0g                  Male life expectancy years
flife           float  %9.0g                  Female life expectancy years
--------------------------------------------------------------------------------
Sorted by:
```

عند إنهاء توصيف المتغيرات يُفترض أن نقوم بحفظ هذه المتغيرات باختيار File > Save أو بكتابة الأمر

```
.save, replace
```

لاحقاً، سوف نقوم بفتح ملف البيانات بالنقر على أيقونة باختيار File > Open أو بكتابة الأمر

```
.use C:\data\Canada1
```

يمكننا الآن التقدم في عملية تحليل البيانات، ربما تلاحظ أن هناك ارتباطاً إيجابياً بين العمر المتوقع للذكور *mlife* والإناث *flife*. وهذان المتغيران يرتبطان بشكل سلبي مع معدل البطالة *unemp*. الارتباط بين العمر المتوقع ومعدل البطالة يكون أقوى للذكور منه للإناث.

```
.correlate unemp mlife flife
(obs=11)
```

	unemp	mlife	flife
unemp	1.0000		
mlife	-0.7440	1.0000	
flife	-0.6173	0.7631	1.0000

ترتيب المشاهدات في ملف البيانات يمكن تغييره باستخدام الأمر **sort**، فمثلاً لإعادة ترتيب المشاهدات من الأصغر للأكبر حسب متغير عدد السكان، قم بطباعة الأمر التالي:

```
.sort pop
```

المتغيرات النصية يمكن ترتيبها أبجدياً، فطباعة الأمر **sort *place*** سوف يعيد ترتيب المشاهدات واضعاً Alberta أولاً، British Columbia ثانياً وهكذا.

الأمر **order** يقوم بترتيب المتغيرات في ملف البيانات، فمثلاً يمكننا وضع معدل البطالة ثانياً، وعدد السكان أخيراً بطباعة الأمر:

```
.order place unemp mlife flife pop
```

محرر البيانات Data Editor توجد به قائمة Tools وبها عدد من الخيارات التي يمكن أن تقوم بنفس المهام السابقة.

يمكننا إجراء تحديد للمتغيرات المراد العمل عليها في نافذة محرر البيانات Data Editor بحيث تعرض فقط المتغيرات التي يُراد تغييرها أو إدخال بيانات بها، فمثلاً الأمر التالي يعرض فقط ثلاثة متغيرات في نافذة محرر البيانات.

```
.edit place mlife flife
```

أو:

```
.edit place unemp if pop> 100
```

الأمر الأخير يستخدم if والتي تعتبر أداة مهمة، وسوف يتم شرحها في الأجزاء اللاحقة.

إنشاء ملف بيانات جديد باستخدام نسخ Copy ولصق Paste :

Creating a New Dataset by Copy and Paste

عندما تكون البيانات الأصلية مخزنة على وسيلة إليكترونية مثل صفحة ويب أو ملف نصي أو ملف إكسيل أو وورد، فيمكننا جلب هذه البيانات لبرنامج ستاتا باستخدام copy وpaste؛ فمثلاً مركز بيانات المناخ الوطني National Climate Data (NCDC) Center يعرض توقعات درجات الحرارة العالمية غير الاعتيادية (الانحرافات عن متوسط درجات الحرارة في الفترة ما بين 1901 إلى 2000 بالدرجات المئوية) لكل شهر في الفترة الماضية حتى يناير 1880، مؤشر NCDC هو أحد المؤشرات التي تعتمد على شبكة عالمية للبيانات من محطات الأرصاد، وقياسات حرارة سطح البحر، ويقوم NCDC بتحديث المؤشر شهرياً، وينشر هذه التحديثات على الإنترنت؛ وأدناه الخمسة أشهر الأولى من هذه البيانات، أول قيمة 0.0623- تشير إلى أن يناير 1880 كان عالمياً °C 0.06 أكثر برودة من متوسط درجات الحرارة في القرن العشرين.

```
1880  1  -0.0623
1880  2  -0.1929
1880  3  -0.1966
1880  4  -0.0912
1880  5  -0.1510
```

بناءً على تفاصيل كيفية تنظيم البيانات الخام (بما فيها القيم المفقودة) قد لايمكننا استخدام نسخ copy البيانات ككل ثم لصقها paste في نافذة محرر

البيانات Data Editor، فقد يتم توسيط خطوة أخرى مفيدة وذلك من خلال وضع فاصلة بين القيم، وأسهل طريقة لعمل ذلك تتم من خلال نسخ جميع الأرقام ثم لصقها في نافذة الملف التنفيذي لبرنامج ستاتا Do-File Editor وهذه النافذة عبارة عن محرر بيانات بسيط يحتوي على العديد من التطبيقات، ثم قم باستخدام وظيفة استبدال في نافذة الملف التنفيذي من خلال اختيار Edit > Find > Replace وذلك لاستبدال كل المسافتين الموجودتين بين البيانات بمسافة واحدة، قم بتكرار هذه الخطوة عدة مرات حتى يتم حذف كل المسافتين وجعلها مسافة واحدة بين كل البيانات، ثم آخر خطوة هي استبدال الكل Replace All المسافة الواحدة بفاصلة؛ بذلك نكون قد قمنا باستخدام Do-File Editor لوضع فواصل بين القيم، وبذلك تكون البيانات في تنسيق معروف لكثير من البرامج، كما يمكننا أيضاً وضع أسماء المتغيرات في الصف الأول مع إضافة فواصل بينها، وتكون البيانات على الشكل التالي:

```
year,month,temp
1880,1,-0.0623
1880,2,-0.1929
1880,3,-0.1966
1880,4,-0.0912
1880,5,-0.1510
```

يمكننا الآن نسخ البيانات من Do-File Editor باختيار Edit > Select All ولصق البيانات في نافذة محرر بيانات خالية في برنامج ستاتا، وذلك باختيار Paste Special مع فاصلة Comma ثم حدد الخيار Treat first row as variable names.

Data Editor (Edit) - [Untitled]

File Edit View Data Tools

year[1] 1880

	year	month	temp
1	1880	1	-.0623
2	1880	2	-.1929
3	1880	3	-.19660001
4	1880	4	-.0912
5	1880	5	-.15099999
6	1880	6	-.1886
7	1880	7	-.1716
8	1880	8	-.1025
9	1880	9	-.1072
10	1880	10	-.18430001

Variable	Label
year	year
month	month
temp	temp

Variables	
Name	year
Label	year
Type	int
Format	%10.0g
Value Label	

Ready Vars: 3 Order: Dataset Obs: 1,584 Filter: Off Mode: Edit

القيام بوضع فواصل بين القيم في ملف امتداده (csv.)، يمكن القيام به باستخدام أي برنامج جداول إلكترونية أو باستخدام برنامج ستاتا نفسه، والذي يجعل البيانات قابلة للنقل بشكل مريح، ولقراءة ملف امتداده csv. باستخدام ستاتا مباشرة يمكن طباعة الأمر **insheet** وذلك كما يلي

```
.insheet using C:\data\global1.csv, comma clear
```

عند فتح البيانات في برنامج ستاتا، يمكننا إضافة توصيف للمتغيرات والبيانات، ثم حفظ النتائج في ملف ستاتا كما يلي:

```
.label data "Global climate"
.label variable year "Year"
.label variable month "month"
.label variable temp "NCDC global temp anomaly
 vs 1901-2000, c"
.save c:\data\global1.dta
.describe
```

```
Contains data from C:\data\global1.dta
  obs:         1,584                        Global climate
 vars:             3                        4 Jul 2012 11:21
 size:        11,088
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
year            int    %8.0g                  Year
month           byte   %8.0g                  Month
temp            float  %9.0g                  NCDC global temp anomaly vs 1901-2000, C
-------------------------------------------------------------------------------
Sorted by:
```

تحديد فئات فرعية من البيانات باستخدام المحددات in و if :

Specifying Subsets of the Data: in and if Qualifiers

العديد من أوامر ستاتا يمكنها تحديد فئة البيانات التي تعمل معها، وذلك بإضافة المحددات **in** أو **if** إلى الأمر نفسه، كما أن هذه المحددات متوافرة في العديد من قوائم ستاتا: ابحث عن if/in أو by/if/in في أعلى نوافذ الحوار. فعلى سبيل المثال، **list in 5** يطلب من ستاتا أن يقوم بعمل قائمة للخمس مشاهدات الأولى، ولعمل قائمة بالخمس مشاهدات الأولى، قم بطباعة الأمر.

```
.list in 1/5
```

	year	month	temp
1.	1880	1	-.0623
2.	1880	2	-.1929
3.	1880	3	-.1966
4.	1880	4	-.0912
5.	1880	5	-.151

الحرف (l) في الأمر أعلاه يحدد آخر مشاهدة، و(10-) يقـوم بتحديـد المشاهدة رقم (10) من أسفل القائمة. فبيانات درجات الحرارة العالمية التـي تغطي فترة 1584 شهرًا تحتوي على أعلى (10) أشهر، والتي كانت فيهـا درجات حرارة غير اعتيادية، وهذا يعني أن درجات الحرارة هذه كانت أعلى من متوسط درجات الحرارة خلال أشهر الفتـرة مـن 1901 إلـى 2000؛ ولإيجاد ذلك يجب علينا أولاً ترتيب درجات الحرارة من أقل قيمة، إلى أعلى قيمة ثم إيجاد الــ (10) درجات من أسفل القائمة كما يلي:

```
.sort temp
.list in -10/l
```

	year	month	temp
1575.	1998	4	.7241
1576.	2003	12	.7317
1577.	2004	11	.7399
1578.	2006	12	.7417
1579.	2010	4	.7515
1580.	2002	3	.7704
1581.	2002	2	.7784
1582.	2010	3	.7802
1583.	1998	2	.8388
1584.	2007	1	.8422

لاحظ أهمية التفرقة بين (1) (رقم واحد، أو أول مشاهدة) وحــرف (l) (حرف إل أو آخر مشاهدة) في الأمر أعلاه.

المحدد **in** يمكن استخدامه بنفس الطريقة مع أغلب أوامر تحليل وتحرير البيانات، ودائماً يشير إلى البيانات كما تم ترتيبها في آخر مرة، أما المحدد **if** فله العديد من التطبيقات، ولكنه يختار المشاهدات بناءً على قيم متغير محدد، فمثلاً لعرض المتوسط الحسابي والانحراف المعياري لدرجة الحرارة غير الاعتيادية قبل سنة 1970 نقوم بطباعة الأمر التالي:

```
.summarize temp if year< 1970
```

Variable	Obs	Mean	Std. Dev.	Min	Max
temp	1080	-.1232613	.1829313	-.7316	.4643

ولتلخيص درجات الحرارة في السنوات الأخيرة نقوم بطباعة الأمر:

```
.summarize temp if year>= 1970
```

Variable	Obs	Mean	Std. Dev.	Min	Max
temp	504	.3159532	.2300395	-.2586	.8422

علامة "<" (أقل من)(*) وعلامة "= >" (أكبر من أو تساوي) هي علامات متعلقة بالأوامر:

==	يساوي
!=	لاتساوي (=~ تقوم بنفس المهمة)
>	أكبر من
<	أصغر من
>=	أكبر من أو تساوي
<=	أصغر من أو تساوي

لاحظ أن علامتي يساوي "==" تشيران إلى الاختبار المنطقي "هل القيمة في الطرف الأيسر تساوي القيمة التي في الطرف الأيمن؟". بالنسبة لبرنامج ستاتا، فإن علامة يساوي واحدة تعني شيئاً آخر مختلفاً وهو: "إجعل القيمة

(*) لاحظ أن اتجاه علامتي أكبر من وأصغر من خاصة باللغة الإنجليزي، وذلك لأن جميع أوامر ستاتا باللغة الإنجليزية، لأن هذه العلامة سوف يكون اتجاهها في الجهة المعاكسة في حالة اللغة العربية.

التي في الطرف الأيسر مساوية للقيمة التي في الطرف الأيمن"، علامة = واحدة ليست علامة متعلقة بالأوامر، ولا يمكن استخدامها مع محدد **if**، وذلك لأن علامة = واحدة لها معان أخرى، فعلامة = واحدة تظهر في تطبيقات محددة مثل الأوزان واختبار الفرضيات.

استخدام علامة أو أكثر متعلقة بالأوامر يمكن أن يتم مع محدد **if** واحد، وذلك باستخدام العلامات المنطقية، والعلامات المنطقية في ستاتا هي:

- & علامة "و" and
- | علامة "أو" or
- ! علامة "ليس" not (~ تقوم بنفس المهمة)

الأقواس تسمح لنا بتحديد أسبقية العمليات، فلعرض ملخص درجات الحرارة غير الاعتيادية لشهري يناير وفبراير للسنوات من 1940 إلى 1969 نقوم باستخدام الأمر التالي:

```
.summarize temp if(month == 1 | month == 2)
 &year>=1940 &year<1970
```

ملاحظة مهمة بشأن البيانات المفقودة: ستاتا سوف يعرض البيانات المفقودة كفترة، ولكن في بعض العمليات (بشكل ملحوظ **sort** و**if** بالرغم من أنها ليست عمليات إحصائية مثل المتوسط الحسابي أو الارتباط)، هي نفس القيم المفقودة والتي تُعامل معاملة أكبر قيم موجبة، فمثلاً بافتراض أننا نريد تحليل بيانات استطلاعات الرأي، الأمر التالي سوف يقوم بعمل جدول انتخاب للمشاركين الذين أعمارهم 65 سنة وأكبر، وكذلك المشاركين الذين أعمارهم عبارة عن قيم مفقودة:

```
.tabulate vote if age>= 65
```

ولكن عند وجود قيم مفقودة، فإننا نحتاج إلى التعامل معها بشكل أكثر تحديداً من خلال استخدام محدد **if** كما يلي:

```
.tabulate vote if age>= 65 & !missing(age)
```

خيار عدم عرض القيم المفقودة، ( )**missing!** يعتبر طريقة عامة لاختيار المشاهدات مع عدم عرض القيم المفقودة، بالرغم من أننا استخدمنا

حتى الآن الرمز الافتراضي للقيم المفقودة وهـو "." إلا أن سـتاتا يمكنـه استخدام أكثر من 27 رمزًا لهذه القيم؛ فاستخدام **if !missing (*age*)** سـوف يضع جانباً القيم المفقودة. وللحصول على معلومات أكثر عن القيم المفقودة، قم باستخدام الأمر **help missing**.

هناك طرق مختلفة لاستبعاد القيم المفقودة من العرض، فدالة **(missing)** تقوم بعرض (1) إذا كانت القيمة مفقودة و(0) إذا لم تكـن مفقـودة، فمـثلاً لوضع جدول الانتخاب للمشاهدات التي ليس لها قيم مفقودة لمتغيرات العمر *age*، والدخل *income* والتعليم *education* قم بطباعة الأمر

```
.tabulate vote if missing (age, income,
 education) ==0
```

أخيراً، وحيث إن القيمة المفقودة الافتراضية هي "." وستاتا يعتبرها أكبر قيمة، وهناك قيم مفقودة أخرى (سوف يتم شرحها لاحقاً) تعتبر أكبر، يمكننا استخدام علامة أقل من أو علامة لايساوي ">" لاستبعاد عرض القيم المفقودة كما يلي:

```
.tabulate vote if age<. &income<. &Education<.
```

المحددان **in** و**if** يمكن استخدامهما لاستبعاد بعض المشاهدات مؤقتاً عند عدم قدرة أمر معين على العمل، هذان المحددان ليس لهما أي تـأثير علـى البيانات المخزنة، فالأمر أدناه سوف يطبق على كل المشاهدات مـالم يـتم استخدام **in** أو **if**، ولحذف متغيرات من البيانات، قم باستخدام الأمـر **drop** (أو استخدم نافذة محرر البيانات Data Editor)، عودة إلى ملف بيانات كنـدا (*Canada1.dta*) يمكننا حذف المتغيرات *mlife* والمتغير *flife* من البيانـات بطباعة الأمر.

```
.drop mlife flife
```

يمكن استخدام كلٍ من **in** أو **if** لتحديد المشاهدات التي يُـراد حـذفها، فمثلاً **drop in 12/13** يعني قم بحذف المشاهدات 12 و13 في البيانات، كما يمكننا كذلك حذف مشاهدات محددة أو متغيرات محددة بالـضغط علـى زر Delete أثناء استخدام نافذة محرر البيانات Data Editor.

بدلاً من حذف مشاهدات أو متغيرات في ملف بيانــات ســتاتا، يمكننـــا أحياناً أن نقوم بتحديد ما نريد الإبقاء عليه؛ فمثلاً بدلاً من استخدام أمر **drop** للمتغير *mlife* والمتغير *flife* من ملف *Canada1.dta* يمكــن القيــام بــنفس المهمة باستخدام الأمر **keep**، وكتابة المتغيرات الثلاثة الأخرى كما يلي:

```
.keep place pop unemp
```

مثل الكثير من التغييرات التي يتم القيام بها في برنامج ستاتا، فإن كــل تغيير لن يتم الإبقاء عليه في ملف البيانات ما لم نقم بحفظ ذلك التغيير، عند هذه النقطة سوف يكون لدينا الخيار (**save** ، **replace**) وبذلك تتم التغييرات في الملف الحالي، كما يمكن حفظ البيانات في ملف جديد باختيــار File > Save As أو بطباعة الأمر **save** ***newname*** ويكون لديك بجهازك ملفــان اثنان للبيانات.

إنشاء واستبدال المتغيرات : Generating and Replacing Variables

أمر الإنشاء **generate** والاستبدال **replace** يتيح لنا إنــشاء متغيــرات جديدة، أو تغيير قيم المتغيرات الحالية، فمثلاً في أغلب المجتمعات الصناعية بكندا، هناك احتمال أن تكون النساء أطول عمراً من الرجال؛ ولتحليل التغير في أعمار الجنسين نقوم بفتح ملف *Canada1.dta*، وإنــشاء متغيــر جديــد يساوي الفرق بين العمر المتوقع للإناث *flife* والعمر المتوقع للذكور *mlife*، وعند استخدام الأمر **generate** أو **replace** سوف نقوم باســتخدام علامـــة يساوي واحدة.

```
.use C:\data\Canada1, clear
.generate gap = flife - mlife
.label variable gap "Female-male life
 expectancy gap"
.describe gap
```

variable name	storage type	display format	value label	variable label
gap	float	%9.0g		Female-male life expectancy gap

```
.list place flife mlife gap
```

	place	flife	mlife	gap
1.	Canada	81.1	75.1	6
2.	Newfoundland	79.8	73.9	5.900002
3.	Prince Edward Island	81.3	74.8	6.5
4.	Nova Scotia	80.4	74.2	6.200005
5.	New Brunswick	80.6	74.8	5.799995
6.	Quebec	81.2	74.5	6.699997
7.	Ontario	81.1	75.5	5.599998
8.	Manitoba	80.8	75	5.800003
9.	Saskatchewan	81.8	75.2	6.600006
10.	Alberta	81.4	75.5	5.900002
11.	British Columbia	81.4	75.8	5.599998
12.	Yukon	80.4	71.3	9.099998
13.	Northwest Territories	78	70.2	7.800003

بالنسبة لولاية Newfoundland فإن القيمة الصحيحة لـ *gap* يفترض أن تكون 79.8 – 73.9 = 5.9 سنة، ولكن بدلاً من ذلك فإن جدول المخرجات أعلاه يعرض القيمة 5.900002 وذلك لأن ستاتا مثله مثل كل برامج الحاسب يقوم بتخزين الأرقام بنظام ثنائي والقيمة 5.9 ليس لها تعبير ثنائي. عدم الدقة البسيطة نتيجة التقريب في النظام الثنائي ليس لها تأثير على النتائج الإحصائية، ولكنها قد تثير الانتباه في قائمة البيانات؛ ويمكننا تغيير تنسيق العرض بحيث يعرض ستاتا قيماً مقرّبة. فالأمر التالي يجعل تنسيق العرض ثابتاً بحيث يكون عرض العمود كافياً لعرض أربعة أرقام وبعد الفاصلة يتم عرض رقم واحد فقط.

```
.format gap %4.1f
```

حتى عند عرض القيمة 5.9 بالأمر أعلاه، فإن الأمر التالي لن يستطيع عرض هذه القيمة عند طلبها تلك القيمة مباشرة.

```
.list if gap == 5.9
```

هذا يحدث لأن ستاتا يعتقد بأن القيمة لا تساوي بالضبط 5.9 (تقنياً ستاتا يخزن قيم *gap* بشكل مفرد كما هي، ولكن كل العمليات الحسابية تتم بالنظام الثنائي، وفي النظام المفرد والثنائي 9.5 لايمكن تحديدها بدقة).

تنسيقات العرض وتنسيقات أسماء المتغيرات والتوصيفات يمكن تغييرها أيضًا وذلك بالنقر مرتين على أي عمود في نافذة محرر البيانات Data Editor، حيث يمكن وضع تنسيقات رقمية ثابتة مثل **%4.1f** كأحد أكثر التنسيقات الرقمية استخداماً، وهناك أيضًا عدة تنسيقات أخرى منها

%*w*.*d*g **تنسيق رقمي عام**، حيث إن *w* يحدد العرض الكلي أو عدد الأعمدة التي سوف يتم عرضها، و*d* أقل عدد من الأرقام التي يجب أن تظهر بعد الفاصلة، أما بخصوص التعبيرات الأسّية (مثل 1.00e+07 والتي تعني 1.00×10^7 أو تعني 10 ملايين) سوف يتم استخدامها بشكل تلقائي عند الحاجة، وذلك لعرض القيم بأفضل شكل ممكن.

%*w*.*d*f **تنسيق رقمي ثابت**، حيث إن *w* يحدد العرض الكلي أو عدد الأعمدة التي يتم عرضها و*d* عدد ثابت من الأرقام التي يجب أن تظهر بعد الفاصلة العشرية.

%*w*.*d*e **تنسيق رقمي تصاعدي**، حيث إن *w* يحدد العرض الكلي أو عدد الأعمدة التي يتم عرضها و*d* عدد ثابت من الأرقام التي يجب أن تظهر بعد الفاصلة العشرية.

فعلى سبيل المثال، إذا نظرنا إلى الجدول (1.2) فإن عدد سكان كندا كان 29,606,100 تقريباً، وكان عدد سكان إقليم يوكون Yukon 30,100 تقريباً، الجدول أدناه يعرض كيف تظهر هذه الأرقام باستخدام تنسيقات مختلفة.

التنسيق	كندا	يوكون Yukon
%9.0g	2.96e+07	30100
%9.1f	29606100.0	30100.0
%12.5e	2.96061e+07	3.01000e+04

بالرغم من أن الأرقام المعروضة تظهر بصور مختلفة، فإن قيمها متطابقة تماماً، وهذه التنسيقات لا تؤثر على العمليات الحسابية. خيارات التنسيق الأخرى للأرقام المعروضة تتضمن استخدام الفواصل، والمحاذاة

لليمين واليسار، وكذلك عدد الأصفار التي تظهر، وهناك أيضًا تنسيقات خاصة للتواريخ ولمتغيرات السلاسل الزمنية والمتغيرات النصية، وللحصول على معلومات أكثر حول التنسيقات قم بطباعة الأمر **help format**.

الأمر **replace** يمكنه أن يقوم بنفس العمليات الحسابية التي يقوم بها **generate** ولكنه يقوم بتغييرقيم المتغير الموجود بدلاً من إنشاء متغير جديد. فعلى سبيل المثال، إذا افترضنا أنه لدينا بيانات الدخل بالدولار، وقررنا أنه من الأفضل العمل مع هذه البيانات بحيث تكون بالآلاف، وللقيام بعملية تحويل قيم الدولار إلى آلاف يمكننا التقسيم على 1000:

```
.replace income = income/1000
```

الأمر **replace** يمكنه القيام بمثل هذه التغييرات بالجملة، أو يمكن استخدامه مع المحددات **in** و**if** لتحرير بيانات معينة فقط، بافتراض أن متغيرات المسح تتضمن العمر *age* وسنة الميلاد *born* فأمر ستاتا مثل الأمر أدناه سوف يقوم بتصحيح واحد أو أكثر من المتغيرات التي تم طباعتها بطريقة خاطئة مثل 299 بدلاً من 29

```
.replace age = 29 if age == 299
```

أو يمكن القيام بنفس الأمر بطريقة أخرى بطباعة الأمر أدناه لتصحيح قيم العمر *age* للمشاهدة رقم 1453

```
.replace age = 29 in 1453
```

وللقيام بذلك هناك مثال أكثر تعقيداً.

```
.replace age = 2012-born if missing(age)| age+1
 < 2012-born
```

الأمر أعلاه يستبدل قيم متغير العمر *age* بقيمة 2012 مطروحاً منها سنة الميلاد *born* إذا كان العمر *age* قيمة مفقودة أو إذا كان العمر المعروض (سنة الميلاد مضافاً إليها 1) أقل من 2012 مطروحاً منها سنة الميلاد.

الأمر **generate** والأمر**replace** يُعتبران أدوات لإنشاء متغيرات تصنيفيّة، فقد لاحظنا سابقاً أن البيانات الكندية تتضمن عدة أنواع من

المشاهدات (حيث إنها تشمل إقليم ينو 10 ولايات ودولة واحدة) وبالرغم من أن المحددين **in** و**if** يسمحان لنا بفصل هذه المشاهدات، فـإن الأمـر **drop** يمكنه إزالة هذه المشاهدات من البيانات. وربما يعتبر أكثر الأمور سهولة هو وضع متغيرات تصنيفيّة توضح نوع المشاهدات؛المثال التالي يوضح طريقة إنشاء مثل هذه المتغيرات باستخدام ملف بيانات كندا *Canada1.dta*، ونبـدأ بإنشاء المتغير *type* كمتغير ثابت يساوي 1 لكل مشاهدة، ثم نقوم باسـتبدال هذه القيمة بالقيمة 2 لإقليم يوكـون Yukon والأقـاليم الـشمالية الغربيـة Northwest Territory والقيمة 3 لكندا، والخطوات الأخيرة تتضمن توصيف المتغير الجديد *type* وتعريف التوصيفات للقيم 1، 2، 3.

```
.use "C:\data\Canada1.dta", clear
.generate type = 1
.replace type = 2 if place == "Yukon" |
place =="northwest Territories"
.replace type = 3 if place == "Canada"
.label variable type "Province, territory or
 nation"
.label define type1b1 1 "Province" 2
 "Territory" 3 "Nation"
.label values typetype1b1
.list
```

	place	pop	unemp	mlife	flife	type
1.	Canada	29606.1	10.6	75.1	81.1	Nation
2.	Newfoundland	575.4	19.6	[illegible].9	79.8	Province
3.	Prince Edward Island	136.1	19.1	74.8	81.3	Province
4.	Nova Scotia	937.8	13.9	74.2	80.4	Province
5.	New Brunswick	760.1	13.8	74.8	80.6	Province
6.	Quebec	7334.2	13.2	74.5	81.2	Province
7.	Ontario	11100.3	9.3	75.5	81.1	Province
8.	Manitoba	1137.5	8.5	75	80.8	Province
9.	Saskatchewan	1015.6	7	75.2	81.8	Province
10.	Alberta	2747	8.4	75.5	81.4	Province
11.	British Columbia	3766	9.8	75.8	81.4	Province
12.	Yukon	30.1	.	71.3	80.4	Territory
13.	Northwest Territories	65.8	.	70.2	78	Province

كما يظهر في الجدول السابق، فإن توصيف القيم لمتغير تصنيفي يتطلب أمرين اثنين، فالأمر **label define** يحدد ماهي التوصيفات التي ترافق الأرقام، والأمر **label values** يحدد ما هو المتغير الذي تنطبق عليه التوصيفات، إحدى مجموعة التوصيفات (والتي تم إنشاؤها بالأمر **label define**) يمكن تطبيقها على أي عدد من المتغيرات الرقمية (وذلك بإضافتها إلى الأمر **label values**)، توصيفات القيم يمكن أن تتضمن 32000 حرف، ولكن من الأفضل ألا تكون توصيفات المتغيرات طويلة جداً.

الأمر **generate** يمكنه إنشاء متغيرات جديدة، والأمر **replace** يمكنه القيام بتخفيض القيم باستخدام أي خليط من المتغيرات القديمة والثوابت والقيم العشوائية والتعبيرات، بالنسبة للمتغيرات الرقمية يجب أن يُطبّق عليها الإشارات الحسابية التالية:

+ جمع
− طرح
* ضرب
/ قسمة
^ الأس

الأقواس سوف تتحكم في ترتيب العمليات الحسابية، وبدون استخدام الأقواس، فإنه سوف يتم اتباع الترتيب المعتاد للعمليات الحسابية، وبالنسبة للعلامات الحسابية، فإن العلامة الوحيدة التي يتم استخدامها مع المتغيرات النصية هي علامة الجمع "+" حيث إنها تقوم بدمج متغيرين نصيين في متغير واحد.

وبالرغم من اختلاف أغراض استخدام الأمر **generate** والأمر **replace** فإنهما متشابهان من حيث التركيبة، حيث إنهما يستخدمان نفس المعاملات الحسابية والمنطقية التي يستخدمها ستاتا، كما أنهما يستخدمان المحددين **in** أو **if**، كما أن هذه الأوامر يمكنها استخدام مجموعة من الدوال الخاصة ببرنامج ستاتا، والتي سوف يتم شرحها لاحقاً.

رموز القيم المفقودة : Missing Value Codes

الأمثلة التي تم شرحها حتى الآن، تتضمن رمزًا واحدًا فقط للقيم المفقودة. والقيمة المفقودة يعتبرها ستاتا أعلى قيمة عند ترتيب القيم من أقل قيمة إلى أكبر قيمة. وبصفة عامة، فإن وجود القيم المفقودة في بعض البيانات ترجع لأسباب عدة؛ ويمكننا استخدام عدة أنواع من الرموز لتمثل القيم المفقودة، وذلك بإضافة امتداد لرموز القيم المفقودة، وهذه القيم المفقودة بامتداد سوف يعتبرها ستاتا قيماً كبيرة عند ترتيب البيانات من أعلى قيمة إلى أقل قيمة، فمثلاً سوف يتم ترتيب القيم المفقودة من ".a" إلى ".z" كما أن رموز القيم المفقودة بامتداد يمكن أن تتم إضافة توصيف لها على عكس الرمز الافتراضي للقيمة المفقودة "." والذي لايمكن إضافة توصيف له.

تظهر الحاجة إلى استخدام أنواع مختلفة من رموز القيم المفقودة في الدراسات الاستقصائية عند استخدام استمارة الاستبيان. فمثلاً قد لا نجد إجابة لسؤال "في أي سنة تزوجت؟" وذلك لأن أحد أفراد العينة لم يسبق له الزواج، ولا يمكنك أن تتجاهل الإجابات الخاصة بهذا السؤال؛ مستخدماً ملف البيانات *Granite2011_6.dta* والذي يحتوي على بيانات من دراسة استقصائية حول وجهات النظر السياسية بولاية نيوهامبشير والذي قام به مركز جرانيت لاستطلاع الرأي في جامعة هامبشير بالولايات المتحدة، ففي أحد أسئلة الاستطلاع، تم سؤال أفراد العينة عن مستوى اهتمامهم بالانتخابات العامة لسنة 2012 (*genint*) وسوف نستخدم هذا السؤال لشرح رموز القيم المفقودة بامتداد.

في البداية يظهر أن مستوى الاهتمام *genint* واضح، ولكن من الصعب القيام ببعض التحليلات الإحصائية مع هذا السؤال:

```
.tabulate genint
```

Interest in 2012 pres. election	Freq.	Percent	Cum.
extremely interested	245	47.48	47.48
very interested	168	32.56	80.04
somewhat interested	72	13.95	93.99
not very interested	28	5.43	99.42
don't know	2	0.39	99.81
no answer	1	0.19	100.00
Total	516	100.00	

يظهر من الجدول أعلاه أن القيم الأربع الأولى تمثل مقياس الاهتمام، وتم توصيفها من "مهتم للغاية" "extremely interested" إلى "غير مهتم جداً" "not very interested"، أما آخر قيمتين فهما "لا أعرف" "don't know" و"لا إجابة" "no answer" وهاتان الإجابتان لايمثلان جزءًا من مقياس الاهتمام، وإنما هما نوع من عدم الإجابة. مركز جرانيت لاستطلاع الرأي يستخدم أرقاماً خاصة في الدراسات الاستقصائية لتمثل عدة أنواع من عدم الإجابة. وفي مثالنا هذا، فإن الرقم 98 يعني أن المشارك في الدراسة قد أجاب بأنه لايعلم مستوى اهتمامه، في حين، أن الرقم 99 يعني لا إجابة تم إعطاؤها للسؤال؛ ويمكننا أن نرى هذه القيم الرقمية إذا قمنا باستخراج نفس الجدول أعلاه بدون توصيف القيم.

```
.tabulate genint, nolabel
```

Interest in 2012 pres. election	Freq.	Percent	Cum.
1	245	47.48	47.48
2	168	32.56	80.04
3	72	13.95	93.99
4	28	5.43	99.42
98	2	0.39	99.81
99	1	0.19	100.00
Total	516	100.00	

أي حسابات إحصائية للمتغير *genint* لن تكون دقيقة، وذلك بسبب وجود الرقمين 98 و99، فمثلاً لمعرفة المتوسط الحسابي للمؤهلات العلمية لأفراد العينة للمتغير *genint*، فإن ذلك سوف يكون عديم النفع، وذلك لأن القيمتين 98 و99 تم تضمينهما في حساب المتوسط الحسابي.

```
.tabulate educ, summarize (genint)
```

Highest degree completed	Summary of Interest in 2012 pres. election Mean	Std. Dev.	Freq.
HS or les	2.8275862	8.9668722	116
Tech/some	3.5	12.451587	120
College g	1.672956	.82290667	159
Postgrad	1.5775862	.80380467	11[illegible]
Total	2.3424658	7.4366697	511

هناك حاجة للحصول على نسخة مطوّرة من الجدول السابق، وسوف نسميها المتغير الجديد *genint2*، والشكل الجديد للجدول سوف يكون مختلفاً عن السابق، وذلك لثلاثة أسباب هي:

أولاً: سوف نعكس القيم من 1 إلى 4، بحيث إن القيم الأعلى تشير إلى اهتمام أكبر بدلاً من اهتمام أقل، وهذا يجعل عملية التفسير منطقية أكثر.

```
.generate genint2 = 5 - genint if genint<90
```

ثانياً: القيم 98 و99 يُفترض اعتبارها قيماً مفقودة، ولن يتم تضمينها في حساب المتوسط الحسابي، والحسابات الإحصائية الأخرى، لذلك سوف نستخدم رمز القيمة المفقودة a. لتمثل "لا أعرف" "don't know" والتي تم تمثيلها سابقاً بالرمز 98، والرمز b. لتمثل "لا إجابة" "no answer" والتي تم تمثيلها سابقاً بالرمز 99.

```
.replace genint2 = .a if genint == 98
.replace genint2 = .b if genint == 99
```

ثالثاً: توصيف القيم يمكن اختصاره بجمل قصيرة، فبدلاً من استخدام "مهتم للغاية" "extremely interested" إلى رقم، وبذلك فإن التوصيف سوف يأخذ حيزًا أقل في الرسوم البيانية والجداول.

```
.label variable genint2 "interest in 2012
 election (new)"
.label define genint2 1 "Not very" 2 "Somewhat"
 3 "Very"
 4 "Extremely" .a "DK" .b "NA"
.label values genint2genint2
```

الخطوة الأخيرة المهمة، وهي استخراج جدول للمتغيرات الجديدة، والمتغيرات القديمة للمقارنة بينهما والتأكد من أن جميع الأوامر قد قامت بما هو مُفترض.

```
.tabulate genint genint2, miss
```

Interest in 2012 pres. election	Interest in 2012 election (new) Not very	Somewhat	Very	Extremely	Total
extremely interested	0	0	0	245	245
very interested	0	0	168	0	168
somewhat interested	0	72	0	0	72
not very interested	28	0	0	0	28
don't know	0	0	0	0	2
no answer	0	0	0	0	1
Total	28	72	168	245	516

Interest in 2012 pres. election	Interest in 2012 election (new) DK	NA	Total
extremely interested	0	0	245
very interested	0	0	168
somewhat interested	0	0	72
not very interested	0	0	28
don't know	2	0	2
no answer	0	1	1
Total	2	1	516

بعد إجراء هذه التعديلات، فإن هذا المتغير أصبح قابلاً للتحليل بشكل أكثر وضوحاً من ذي قبل، فمثلاً أصبح من السهل أن نحدد المتوسط الحسابي لمستوى الاهتمام بالانتخابات مع تحديد المستوى التعليمي لأفراد العينة.

```
. tabulate educ, summ(genint2)
```

Highest degree completed	Summary of Interest in 2012 election (new) Mean	Std. Dev.	Freq.
HS or les	3	.98229949	115
Tech/some	3.1101695	.89427331	118
College g	3.327044	.82290667	159
Postgrad	3.4224138	.80380467	116
Total	3.2244094	.88647221	508

عند التعامل مع أرقام محددة (مثل 98 و 99 في المثال أعلاه) والتي تشير إلى القيم المفقودة، فإنه من الأفضل تغيير رموز القيم المفقودة بالطريقة السابقة حتى لا يقوم ستاتا بإدخالها ضمن أي حسابات إحصائية، يمكن

بسهولة القيام بذلك لجميع المتغيرات، وذلك باستخدام الأمر **mvdecode** فمــثلاً يمكن إعطاء الأمر.

```
.mvdecode genintincomeage,mv(97=.\98=.a \99=.b)
```

الأمر أعلاه سوف يغير أي قيم للمتغيرات *genint*، *income*، *age* مــن 97 إلى "." ومن 98 إلى "a." وهكذا، القيم من a. و b. (حتى القيمة z.) قيم مفقودة يمكن إدخال توصيف لها، ولكن لايمكن إدخال توصيف لرمز القيمــة المفقودة ".". وكما هو معتاد، فإن هذه التعديلات التي تم القيام بها لاتــصبح دائمة حتى يتم حفظ البيانات، ومن الأفضل حفظها في ملــف باســم جديــد كإجراء احتياطي، فربما قد نحتاج إلى العودة إلى البيانات الأصلية مــستقبلاً لأي سبب من الأسباب.

استخدام الدوال : Using Functions

هذا الجزء يعرض قائمة بالعديد من الدوال المتوافرة للاستخدام مع الأمر **generate** والأمر **replace**، فعلى سبيل المثال، يمكننا إنشاء متغيــر جديــد باسم *loginc*، وهو يساوي اللوغــاريتم الطبيعــي للمتغيــر *income* وذلــك باستخدام دالة اللوغاريتم الطبيعي **ln** مع الأمر **generate** كما يلي:

```
.generate loginc = ln(income)
```

دالة **ln** هي واحدة من الدوال الرياضية في برنــامج ســتاتا، الأمثلــة الأخرى تتضمن **log10(*x*)** للوغاريتم طبيعــي أساســه 10، **int(*x*)** للعــدد الصحيح *x*، **exp(*x*)** وهي عبارة عن *e* مرفوعة للأس *x* (*e* تساوي تقريبــاً 2.718281828) وهناك العديد من الدوال الأخرى. وللحصول على قائمــة كاملة بتفاصيل هذه الدوال، قم بطباعة الأمر **help math functions**.

وهناك أيضًا العديد من دوال الكثافة الاحتمالية، ويمكنك الحصول على قائمة كاملة بهذه الدوال بطباعة الأمر **help density functions** أو من خلال الاطلاع على دليل المستخدم لبرنامج ستاتا، حيث تحتوي قوائم هذه الــدوال على تعريفات لهذه الدوال، وتركيب معاملاتها، وكيفية تعاملهــا مــع القــيم المفقودة. فعلى سبيل المثال، دالة **invnormal(*p*)** تعطي التوزيــع الطبيعــي

المعياري التراكمي أو قيمة z المرتبطة بدرجة الاحتمال p، وتتضمن الدوال الأخرى قيمة بيتا، والتوزيع الثنائي، ومربع كاي، واختبار t، واختبار F، وتوزيع جاما، والتوزيعات المنتظمة، وهناك دالة أخرى مهمة جداً خاصة بالمحاكاة وهي **runiform()** وتستخدم لإنشاء الأرقام شبه العشوائية وذلك لاستخراج القيم من التوزيعات المنتظمة، وهذه القيم نظرياً تكون في نطاق بين 0 و1 تقريباً وتكتب على الشكل [0,1).

برنامج ستاتا يعرض العديد من دوال التواريخ، وكذلك التواريخ التي تتعلق بدوال السلاسل الزمنية، وخاصة تلك التي لها تنسيقات خاصة في العرض أو المتغيرات التي تتعلق بالتواريخ؛ ويمكن الحصول على قائمة بتفاصيل تلك الدوال من دليل المستخدم الخاص ببرنامج ستاتا أو بطباعة الأمر **help date functions**؛ دوال التاريخ في العادة تتضمن تواريخ ماضية، والتي تشير إلى عدد الأيام منذ 1 يناير 1960م.

بيانات درجة الحرارة العالمية، والتي تم الإشارة إليها سابقاً في هذا الفصل، تعتبر كمثال للتواريخ الماضية، حيث يحتوي ملف البيانات على بيانات السنة *year* والشهر *month* ولكن لا يوجد متغير واحد يتضمن بيانات الشهر والسنة معاً كمقياس للزمن.

```
.use C:\data\global1.dta, clear
.describe

Contains data from C:\data\global1.dta
  obs:         1,584                          Global climate
 vars:             3                          4 Jul 2012 11:21
 size:        11,088
------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
year            int    %8.0g                  Year
month           byte   %8.0g                  Month
temp            float  %9.0g                  NCDC global temp anomaly vs 1901-2000, C
------------------------------------------------------------------------------
Sorted by:
```

يمكننا إنشاء متغير جديد للتاريخ الماضي يسمى *edate* باستخدام دالة **mdy** وهي عبارة عن اختصار للحروف الأولى للكلمات شهر ويوم

وسنة (month, day, year)، تم احتساب متوسط الحرارة الشهري لبيانات درجات الحرارة العالمية في المثال أعلاه، لذلك فمن الممكن استخدام اليوم 15 من كل شهر (المدخل البديل هو استخدام بيانات شهرية – انظر إلى شرح بيانات الملف *Climate.dta* في الفصل 12) وحيث إن *edate* يمثل عدد الأيام منذ 1 يناير 1960، لذا فإن عدد الأيام قبل هذا التاريخ سوف تظهر بإشارة سالبة.

```
.generate edate = mdy(month, 15, year)
.label variable edate "elapsed date"
.list in 1/5
```

	year	month	temp	edate
1.	1880	1	-.0623	-29205
2.	1880	2	-.1929	-29174
3.	1880	3	-.1966	-29145
4.	1880	4	-.0912	-29114
5.	1880	5	-.151	-29084

ويمكن أن تكون النتيجة أكثر وضوحاً إذا قمنا بتنسيق بيانات المتغير *edate* كمتغير يمثل تاريخ (%td) يعرض الشهر (m) والقرن (C) والسنة (Y) ثم نقوم بتوصيف القيم الرقمية للمتغير *edate* ونعطيها وصف "Jan 1880".

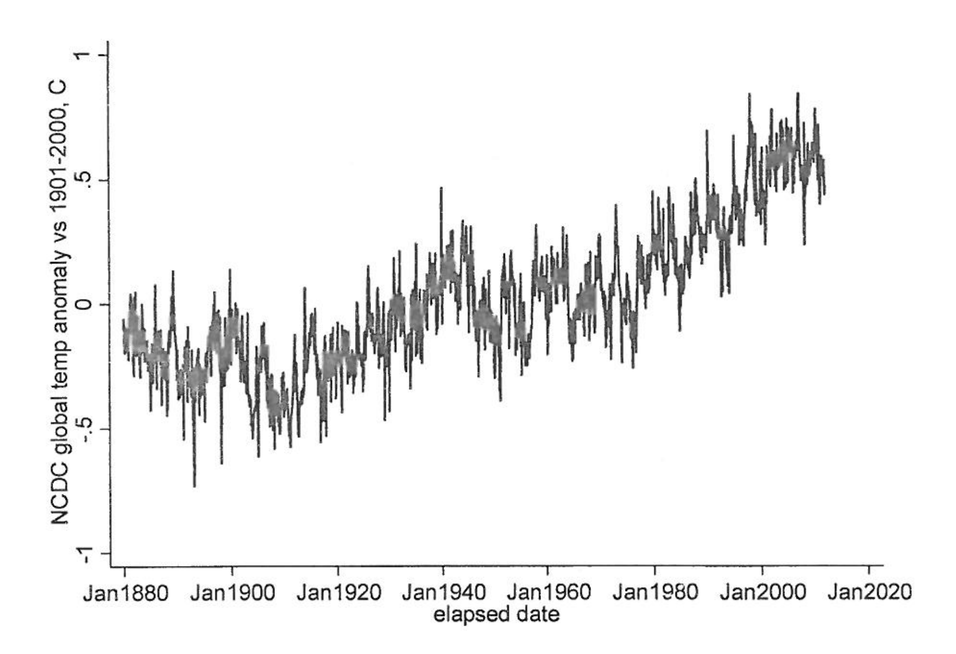
```
.format edate %tdmCY
.list in 1/5
```

	year	month	temp	edate
1.	1880	1	-.0623	Jan1880
2.	1880	2	-.1929	Feb1880
3.	1880	3	-.1966	Mar1880
4.	1880	4	-.0912	Apr1880
5.	1880	5	-.151	May1880

أخيراً نقوم بحفظ البيانات **save** مع المتغير الجديد، وبإنشاء رسم بياني لدرجة حرارة العالم السنوية *temp* مع المتغير *edate* يمكننا الحصول على فكرة عن التغير خلال الفترة الزمنية.

```
.sort yearmonth
```

```
.order year month edate
.save C:\data\glbal12.dta, replace
.graph twoway line temp edate
```



الشكل (1.2)

الأنواع الأخرى من الدوال تتضمن دوال المصفوفات، ودوال الأرقام العشوائية، ودوال النصوص، ودوال السلاسل الزمنية، ودوال البرمجة؛ قم بطباعة الأمر **help** متبوعاً بأي نوع من أنواع الدوال لمشاهدة تفاصيل كل نوع من أنواع الدوال، كما يقوم دليل المستخدم لبرنامج ستاتا بعرض أمثلة أكثر تفصيلاً عن هذه الدوال.

دوال متعددة وعوامل أخرى يمكن استخدامها مع أي أمر حسب الحاجة، والدوال والعوامل الرياضية الأخرى التي سبق شرحها سابقاً يمكن أيضًا استخدامها بطريقة أخرى بحيث لا تقوم بإنشاء أو تعديل أي متغيرات، الأمر **display** يقوم بحساب وعرض النتائج على الشاشة، فمثلاً:

```
.display 2+3
5
```

```
.display log10(10^83)
83
.display invttail(120, .025) * 34.1/sqrt(975)
2.1622305
```

لذا فإن الأمر **display** يمكن استخدامه لعرض الحسابات الإحصائية على الشاشة على خلاف الأوامر **generate** و**replace** والتي تقوم بإجراء تغييرات مباشرة في النتائج الإحصائية. ولتوضيح ذلك، نعود لبيانات الجليد في القطب الشمالي، والتي تم شرحها في الفصل الأول من هذا الكتاب *Arctic9.dta*، أحد المتغيرات *extent* والذي يمثل متوسط المنطقة المغطاة بالجليد بنسبة 15% على الأقل خلال شهر سبتمبر في كل سنة (تم إظهارها في رسم بياني بالشكل 1.1)، فخلال 33 سنة من مشاهدات الأقمار الصناعية، متوسط المنطقة المغطاة بالجليد كان نحو 6.52 مليون كيلومتر مربع.

```
.summarize extent
```

Variable	Obs	Mean	Std. Dev.	Min	Max
extent	33	6.51697	.9691796	4.3	7.88

بعد الحصول على الملخص أعلاه، فإن ستاتا يقوم بحفظ المتوسط الحسابي كمتغير كمي باسم r(mean)

```
.display r(mean)
6.5169697
```

يمكننا استخدام هذه النتيجة لإنشاء متغير جديد باسم *extent0* والذي يُعرّف بأنه الانحراف عن متوسط الفترة 1979 – 2011، *extent0* سوف يكون له نفس الانحراف المعياري الخاص بالمتغير *extent* ولكن المتوسط الحسابي له صفر تقريباً، وهذا يعكس قيمة الانحراف عن المتوسط لكل شهر سبتمبر في كل سنة.

```
.gen extent0 = extent - r(mean)
.summ extent extent0
```

Variable	Obs	Mean	Std. Dev.	Min	Max
extent	33	6.51697	.9691796	4.3	7.88
extent0	33	1.17e-08	.9691796	-2.216969	1.36303

بعد القيام بكل عملية تحليل يقوم برنامج ستاتا بحفظ هذه النتائج مؤقتاً مثل حفظ r(mean) بعد الأمر **summarize** هذه الحسابات مهمة جداً للتحليلات الإحصائية أو البرمجة اللاحقة، ولمشاهدة قائمة كاملة بالأسماء والقيم التي تم حفظها قم بطباعة الأمر **return list**، ففي المثال أعلاه القيم المخزنة r(N) و r(sum_w) و r(mean) تمثل آخر القيم التي قام ستاتا بحفظها في ذاكرته المؤقتة للمتغير *extent*.

```
.return list

          scalars:
                              r(N)  =  33
                          r(sum_w)  =  33
                           r(mean)  =  1.17403088194e-08
                            r(Var)  =  .9393091848549505
                             r(sd)  =  .9691796452954171
                            r(min)  =  -2.21696949005127
                            r(max)  =  1.363030433654785
                            r(sum)  =  3.87430191040e-07
```

برنامج ستاتا يوفر كذلك أوامر أخرى لإنشاء المتغيرات، فالأمر egen (وهو امتداد للأمر **generate**) له مجموعة من الدوال التي تقوم بعدة مهام بطريقة أسرع وأسهل من الأمر **generate**، وهذه المهام تتضمن إنشاء متغيرات جديدة من المجاميع، وأعلى قيمة، وأصغر قيمة، ونطاق الربيعات، والقيم المعلمية والترتيبات أو المتوسطات المرجحة للمتغيرات. فعلى سبيل المثال، لإنشاء متغير جديد باسم *zscore* والذي يساوي القيم المعلمية (متوسط 0، انحراف 1) للمتغير *x* نعطي الأمر التالي:

```
.egen zscore = std(x)
```

أو يمكن القيام بإنشاء متغير جديد *avg* وهو يساوي المتوسط الحسابي لصف قيم كل مشاهدة *x, y, z, w* ويتجاهل أي قيم مفقودة.

```
.egen avg = rowmean (x,y,z,w)
```

لإنشاء متغير جديد باسم *total* يساوي مجموع قيم الصفوف لكل مشاهدة لـ *x, y, z, w* ويعتبر القيم المفقودة صفر.

```
.egen total = rowtotal (x, y, z, w)
```

الأمر التالي يقوم بإنشاء متغير جديد يسمى *xrank*، والذي يحتوي على ترتيب قيم المتغير *x*، علماً بأن $xrank = 1$ لأعلى قيمة في *x*، و $xrank = 2$ لثاني أعلى قيمة، وهكذا.

```
.egen xrank = rank(x)
```

للحصول على قائمة بدوال الأمر **egen** قم باستخدام الأمـــر **help egen** أو قم بالبحث في دليل مستخدم برنامج ستاتا للحصول على أمثلة أكثر عن كيفية استخدام الدالة.

التحويل بين التنسيقات الرقمية والنصّية :

Converting Between Numeric and String Formats

الملف *Canada2.dta* يحتوي على متغير نصي واحد يسمى *place*، كما يحتوي على متغير تصنيفي وصفي *type*، وكلا المتغيرين يحتويان على قيم نصية.

```
.use C:\data\Canada2, clear
.list place type
```

	place	type
1.	Canada	Nation
2.	Newfoundland	Province
3.	Prince Edward Island	Province
4.	Nova Scotia	Province
5.	New Brunswick	Province
6.	Quebec	Province
7.	Ontario	Province
8.	Manitoba	Province
9.	Saskatchewan	Province
10.	Alberta	Province
11.	British Columbia	Province
12.	Yukon	Territory
13.	Northwest Territories	Territory

تحت التوصيف يبقى المتغير *type* متغيراً رقمياً ويظهر بخط أزرق في نافذة محرر البيانات Data Editor أو نافذة Browser وبالضغط على أي خلية

سوف نرى الأرقام، ويمكننا رؤية قائمة بهذه الأرقام باستخدام الأمر **nolabel** كما يلي:

```
.list place type, nolabel
```

	place	type
1.	Canada	3
2.	Newfoundland	1
3.	Prince Edward Island	1
4.	Nova Scotia	1
5.	New Brunswick	1
6.	Quebec	1
7.	Ontario	1
8.	Manitoba	1
9.	Saskatchewan	1
10.	Alberta	1
11.	British Columbia	1
12.	Yukon	2
13.	Northwest Territories	2

المتغيرات النصية والرقمية التي لها توصيفات تسلك سلوكاً مختلفاً عند التحليل. حيث إن أغلب العمليات الإحصائية والعلاقات الرياضية لايمكن استخدامها مع المتغيرات النصية، لذلك فإننا قد نحتاج إلى الحصول على نسخة نصية ورقمية بتوصيفات لبعض المعلومات في البيانات الموجودة لدينا، الأمر **encode** يقوم بإنشاء متغير رقمي بوصف من متغير نصي، الرقم 1 يتم إعطاؤه للحرف الأبجدي الأول للمتغير النصي ورقم 2 للحرف الأبجدي الثاني وهكذا، المثال التالي يقوم بإنشاء متغير رقمي وصفي يسمى *placenum* من المتغير النصي *place* كما يلي:

```
.encode place, gen(placenum)
```

كما أنه من الممكن القيام بالتحويل العكسي، فالأمر **decode** يقوم بإنشاء متغير نصي باستخدام قيم المتغير الرقمي، فمثلاً يمكننا إنشاء المتغير النصي *typestr* من المتغير الرقمي *type*.

```
.decode type, gen(typestr)
```

وعند استخراج قائمة بـالمتغيرات، فـإن المتغيـر الرقمـي الجديـد *placenum*، والمتغير النصي الجديـد *typestr* يتـشابهان مـع المتغيـرات الأصلية.

```
.list place placenum type typestr
```

	place	placenum	type	typestr
1.	Canada	Canada	Nation	Nation
2.	Newfoundland	Newfoundland	Province	Province
3.	Prince Edward Island	Prince Edward Island	Province	Province
4.	Nova Scotia	Nova Scotia	Province	Province
5.	New Brunswick	New Brunswick	Province	Province
6.	Quebec	Quebec	Province	Province
7.	Ontario	Ontario	Province	Province
8.	Manitoba	Manitoba	Province	Province
9.	Saskatchewan	Saskatchewan	Province	Province
10.	Alberta	Alberta	Province	Province
11.	British Columbia	British Columbia	Province	Province
12.	Yukon	Yukon	Territory	Territory
13.	Northwest Territories	Northwest Territories	Territory	Territory

ولكن عند استخدام خيار **nolabel**، فإن الاختلافات تصبح أكثر وضوحاً. حيث إن ستاتا يعتبر المتغير *placenum* والمتغير *type* أرقاماً.

```
.list place placenum type typestr, nolabel
```

	place	placenum	type	typestr
1.	Canada	3	3	Nation
2.	Newfoundland	6	1	Province
3.	Prince Edward Island	10	1	Province
4.	Nova Scotia	8	1	Province
5.	New Brunswick	5	1	Province
6.	Quebec	11	1	Province
7.	Ontario	9	1	Province
8.	Manitoba	4	1	Province
9.	Saskatchewan	12	1	Province
10.	Alberta	1	1	Province
11.	British Columbia	2	1	Province
12.	Yukon	13	2	Territory
13.	Northwest Territories	7	2	Territory

أغلب التحليلات الإحصائية مثل المتوسط الحسابي، والانحرافات المعيارية يمكن القيام بها مع المتغيرات الرقمية فقط، لذلك فإن توصيف هذه المتغيرات لايعتبر ذا أهمية أثناء إجراء الحسابات الإحصائية.

```
.summarize place placenum type typestr

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       place |         0
    placenum |        13           7     3.89444          1         13
        type |        13    1.307692    .6304252          1          3
     typestr |         0
```

أحياناً قد نجد أن قيم أحد المتغيرات النصية في أغلبها أرقاماً، ولتحويل هذه القيم النصية إلى نظيراتها الرقمية، يجب علينا استخدام دالة **real**، فمثلاً في البيانات أدناه، المتغير النصي *siblings* يحتوي على قيمة نصية واحدة "4 or more" والتي يمكن تمثيلها أيضاً برقم.

```
.describe siblings

              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------
siblings        str9   %9s                    Number of siblings (string)
```

```
.list

     +-----------+
     |  siblings |
     |-----------|
  1. |         1 |
  2. |         3 |
  3. |         0 |
  4. |         2 |
  5. | 4 or more |
     +-----------+
```

```
.generate sibnum = real(siblings)
```

المتغير الجديد *sibnum* أصبح الآن متغيراً رقمياً مع قيمة مفقودة واحدة وهي "4 or more"

```
.list
```

	siblings	sibnum
1.	1	1
2.	3	3
3.	0	0
4.	2	2
5.	4 or more	.

الأمر **destring** يوفر طريقة أكثر مرونة لتحويل المتغيرات النصية إلى متغيرات رقمية، ففي المثال أعلاه، يمكننا إجراء نفس المهام، وذلك بطباعة الأمر التالي:

```
.destring siblings, generate(sibnum) force
```

للحصول على معلومات أكثر حول تركيبة هذا الأمر وخياراته، اطبع الأمر **help destring**.

إنشاء متغيرات تصنيفية وترتيبية جديدة :

Creating New Categorical and Ordinal Variables

في الجزء السابق، تم شرح كيفية إنشاء متغير تصنيفي *type*، وذلك للتفرقة بين الأقاليم (دولة أو مقاطعات) (province and nation) في ملف البيانات الكندية، ويمكننا أن نقوم بإنشاء متغيرات تصنيفية وترتيبية بطرق عديدة، هذا الجزء يعطي بعض الأمثلة عن ذلك.

المتغير *type* يحتوي على ثلاثة تصنيفات:

```
.tabulate type
```

Province, territory or nation	Freq.	Percent	Cum.
Province	10	76.92	76.92
Territory	2	15.38	92.31
Nation	1	7.69	100.00
Total	13	100.00	

بافتراض أننا نريد إعادة كتابة المتغير *type* ليكون عبارة عن تعبير ثنائي أو وهمي ممثل بالقيمتين 0 أو 1، الأمر **tabulate** سوف يقوم بإنشاء متغيرات وهمية بطريقة آلية إذا قمنا بإضافة خيار **generate** إليه. في المثال التالي هناك نتائج لمجموعة من المتغيرات *type1*، *type2*، *type3* وكل متغير منها يمثل تصنيفاً واحدًا من ثلاثة تصنيفات من المتغير *type*.

.tabulate *type*, generate(*type*)

Province, territory or nation	Freq.	Percent	Cum.
Province	10	76.92	76.92
Territory	2	15.38	92.31
Nation	1	7.69	100.00
Total	13	100.00	

.describe

```
Contains data from c:\data\Canada2.dta
  obs:            13                          Canadian dataset 2
 vars:            11                          18 Apr 2013 19:57
 size:           741
--------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
--------------------------------------------------------------------------------
place           str21  %21s                   Place name
pop             float  %9.0g                  Population in 1000s, 1995
unemp           float  %9.0g                  % 15+ population unemployed, 1995
mlife           float  %9.0g                  Male life expectancy years
flife           float  %9.0g                  Female life expectancy years
type            float  %9.0g       typelbl    Province, territory or nation
placenum        long   %21.0g      placenum   Place name
typestr         str9   %9s                    Province, territory or nation
type1           byte   %8.0g                  type==Province
type2           byte   %8.0g                  type==Territory
type3           byte   %8.0g                  type==Nation
--------------------------------------------------------------------------------
Sorted by:
     Note:  dataset has changed since last saved
```

.list *place type type1-type3*

	place	type	type1	type2	type3
1.	Canada	Nation	0	0	1
2.	Newfoundland	Province	1	0	0
3.	Prince Edward Island	Province	1	0	0
4.	Nova Scotia	Province	1	0	0
5.	New Brunswick	Province	1	0	0
6.	Quebec	Province	1	0	0
7.	Ontario	Province	1	0	0
8.	Manitoba	Province	1	0	0
9.	Saskatchewan	Province	1	0	0
10.	Alberta	Province	1	0	0
11.	British Columbia	Province	1	0	0
12.	Yukon	Territory	0	1	0
13.	Northwest Territories	Territory	0	1	0

إعادة كتابة المعلومات التصنيفية كمجموعة من المتغيرات الوهمية، لايتضمن فقداناً لأي معلومات. ففي المثال أعلاه، نجد أن المتغيرات من *type1* إلى المتغير *type3* معاً توضح نفس المعلومات التي يوضحها المتغير *type* نفسه، وأحياناً يختار المحللون إعادة كتابة المتغيرات القابلة للقياس في شكل تصنيفي أو ترتيبي حتى عند فقد النتائج جزءًا كبيرًا من المعلومات. فعلى سبيل المثال، المتغير *unemp* في الملف *Canada2.dta* يعطي قياساً لمعدل البطالة واستبعاد كندا من البيانات سوف يؤدي إلى أن يكون معدل البطالة في المدى ما بين 7% و19.6% مع متوسط حسابي 12.26.

```
.summarize unemp if type !=3
```

Variable	Obs	Mean	Std. Dev.	Min	Max
unemp	10	12.26	4.44877	7	19.6

عند هذه النقطة إدخال كندا ضمن التحليل يؤدي إلى عدم توافق البيانات، لذلك سوف نقوم باستبعادها بالأمر التالي:

```
.drop if type==3
```

هناك نوعان من الأوامر يمكن استخدامهما لإنشاء متغير وهمي يسمى *unemp2* يساوي 0 إذا كان معدل البطالة أقل من المتوسط (12.26) ويساوي

1 إذا كان معدل البطالة أكبر من أو يساوي المتوسط، وقيمة مفقودة عند وجود أي قيمة مفقودة ضمن بيانات المتغير *unemp*، أما الأمر الثاني فهو يُسمى أمراً ترتيبياً، وهذا الأمر يقوم باعتبار القيم المفقودة أعلى القيم الموجودة ضمن البيانات.

```
.generate unemp2 = 0 if unemp< 12.26
 (7missing values generated)
.replace unemp2 = 1 if unemp>= 12.26 & !missing(unemp)
 (5real changes made)
```

ربما نحتاج إلى وضع القيم ضمن مجموعات لتكوّن متغير قياس، وهذا يعني أنه يجب علينا القيام بإنشاء متغير تصنيفي أو ترتيبي. ويمكن القيام بذلك باستخدام الدالة **autocode** (انظر استخدام الدوال) والتي تقوم بوضع متغيرات القياس ضمن مجموعات بشكل تلقائي؛ ولإنشاء متغير ترتيبي جديد يسمى *unemp3* والذي يصنف قيم المتغير *unemp* ضمن ثلاث مجموعات متساوية العرض ويكون الفراغ بينها من 5 إلى 20 نقوم بطباعة الأمر:

```
.generate unemp3 = autocode(unemp,3,5,20)
 (2missing values generated)
```

يمكن إنشاء قائمة تعرض بيانات المتغير الوهمي الجديد (*unemp2*) والمتغير الترتيبي (*unemp3*) والمتعلقة بقيم بمتغير القياس الأصلي *unemp*.

```
.list place unemp unemp2 unemp3
```

	place	unemp	unemp2	unemp3
1.	Newfoundland	19.6	1	20
2.	Prince Edward Island	19.1	1	20
3.	Nova Scotia	13.9	1	15
4.	New Brunswick	13.8	1	15
5.	Quebec	13.2	1	15
6.	Ontario	9.3	0	10
7.	Manitoba	8.5	0	10
8.	Saskatchewan	7	0	10
9.	Alberta	8.4	0	10
10.	British Columbia	9.8	0	10
11.	Yukon	.	.	.
12.	Northwest Territories	.	.	.

استخدام المخفّضات الصريحة مع المتغيرات :

Using Explicit Subscripts with Variables

عندما تكون هناك بيانات في ذاكرة برنامج ستاتا، فإن هـذه البيانـات تُستخدم لتعريف متغيرات نظامية محددة، وهذه المتغيرات تقوم بوصف تلك البيانات، فمثلاً N_ تمثل مجموع عدد المـشاهدات بينمـا n_ تمثـل عـدد المشاهدات، حيث إن n=1_ للمشاهدة الأولى، n=2_ للمشاهدة الثانية وهكـذا حتى المشاهدة الأخيرة (n=_N_)، وإذا قمنا بطباعة الأمر أدناه، فإنه يقـوم بإنشاء متغير جديد يسمى caseID وهو عبارة عن رقم كل مشاهدة كما هـي في ترتيبها الحالي.

```
.generate caseID = _n
```

ترتيب البيانات هو طريقة لتغيير رقم قيمة كل مشاهدة n_ ، ولكن قيمتها في المتغير الجديد *caseID* سوف تبقى بدون تغيير، لذلك إذا قمنـا بترتيـب البيانات بطريقة أخرى، فإنه يمكننا أن نعود إلى ترتيبها السابق بطباعة الأمر

```
.sort caseID
```

إنشاء وحفظ الأرقام غير المتكررة المحددة لكل مشاهدة والتـي تقـوم بترتيب المشاهدات في مراحل مبكرة أثناء العمل على البيانـات يمكـن أن تساعد لاحقاً في إدارة هذه البيانات.

يمكننا أن نستخدم المخفّضات الصريحة مع أسماء المتغيرات لتحديد رقم مشاهدة معينة، فعلى سبيل المثال، لعرض المشاهدة الرابعـة فـي بيانـات درجات الحرارة العالمية *global2.dta* والتي كانت في أبريل 1880 وكـان الانحراف في درجة الحرارة 0.0912 °C- نقوم بطباعة الأمر:

```
.display temp[4]
 -.0912
```

وبالمثل، فإن *temp*[5] تعرض المشاهدة الخامسة التي تمثل الانحـراف في درجة الحرارة في مايو 1880 وهي 151 °C.-

```
.display temp[5]
 -.15099999
```

المخفضات الصريحة، والمتغير النظامي n_ لهما علاقة خاصة عندما تكون البيانات متسلسلة، ففي مثال درجات الحرارة أعلاه بالنسبة للمتغير *temp* فإن *temp*[_n] تشير إلى المشاهدة رقم nth_ أما *temp*[_n-1] فإنه يشير إلى درجة الحرارة السابقة و*temp*[_n+1] يشير إلى درجة الحرارة اللاحقة، لذلك فإننا قد نحتاج إلى إنشاء متغير جديد باسم *diftemp* وهو يساوي التغير في *temp* منذ الشهر الماضي.

```
.generate diftemp = temp - temp[_n-1]
```

الفصل (12) سيدور حول تحليل السلاسل الزمنية، ويشرح بالتفصيل هذا الموضوع.

استيراد بيانات من برامج أخرى :

Importing Data from Other Programs

الجزء السابق قام بشرح كيفية إدخال البيانات وتحريرها في نافذة محرر البيانات Data Editor، ولكن إذا كانت البيانات مخزنة ومنسقة في ملف جداول إلكترونية، يمكننا فقط نسخ ولصق البيانات في محرر بيانات خال، أو يمكن لستاتا القيام باستيراد هذه البيانات من ملف إكسيل مباشرة من خلال القائمة كما يلي:

- File > Import > Excel spreadsheet (*.xls; *.xlsx)

أو يمكن استخدام الأمر **import excel**، وببساطة يمكننا استيراد الورقة الأولى في ملف إكسيل المسمى *snowfall.xls* بطباعة الأمر

```
.import excel using C:\data\snowfall.xls, clear
```

ولكن الجداول الإلكترونية تحتوي على عناوين وملاحظات، وعناوين فرعية، وورقات متعددة، ورسومات بيانية، وخصائص أخرى تعقد عملية استيراد البيانات، ولتقييد عملية استيراد البيانات **import** لنطاق خاص من الخلايا نقوم باستخدام الخيار **cellrange()** كما أن الخيار **sheet()** يمكنه تحديد الورقة التي تريد استيراد البيانات منها في ملف إكسيل، والخيار

firstrow يحدد لبرنامج ستاتا أن خلايا الصف الأول تحتوي على أسماء المتغيرات، فمثلاً في الملف إكسيل *snowfall.xls* الورقة الأولى اسمها "Berlin" تحتوي على بيانات تاريخية عن سقوط الثلوج لقرية برلين Berlin بولاية هامبشير بالولايات المتحدة والتي تم مناقشتها في بحث Hamilton et al. (2003) والبيانات موجودة في النطاق من الخلية A5 إلى الخلية O56، والعمود 4 يحتوي على أسماء المتغيرات.

```
.import  excel  using  C:\data\snowfall.xls,
 sheet("Berlin")
    cellrange(a4:o56) firstrow clear
```

بالرغم من أن خاصية استيراد بيانات إكسيل **import excel** تعتبر دقيقة نوعاً ما، فإن إعداد البيانات وتنسيقها في ورقة إكسيل يجعل العملية أكثر سرعة، ويسهّل على ستاتا عملية تحليل البيانات، فمثلاً إذا كان هناك بعض أسماء المتغيرات في ورقة إكسيل يجب أن تقابل معايير ستاتا، وهذه المعايير مثلاً يجب أن تبدأ بحرف أو شرطة تحتيّة " _ " ويجب ألا تحتوي على فراغات، أما القيم المفقودة فيجب أن يتم استبدالها بفراغات أو رموز رقمية، والحروف النصية يجب حذفها من الخلايا بالأعمدة وتمثيلها بمتغيرات رقمية.

يقوم ستاتا بشكل تلقائي بتحديد ما إذا كانت بيانات كل عمود تمثل متغيراً رقمياً أو نصياً، وإذا كانت هناك قيم غير رقمية في أي عمود، فإن ستاتا سوف يعتبر ذلك العمود يحتوي على بيانات متغير نصي، وهذا يعني أن الحسابات الإحصائية مثل المتوسط الحسابي، والارتباط لن تكون ممكنة مع هذا المتغير النصي، وإذا كانت أغلب القيم هي قيم رقمية، فيمكننا إنشاء متغير رقمي جديد (وجعل قيمه النصية كرموز للقيم المفقودة) باستخدام دالة real()

```
.generate newvar = real(oldvar)
```

وبالمثل، فإننا قد نحتاج إلى القيام بهذه العملية عند نسخ ولصق بيانات في محرر البيانات Data Editor وقبل اختيار البيانات المراد نسخها، فإننا قد نحتاج إلى تحرير ملف إكسيل الذي يحتوي على هذه البيانات. إحدى الطرق السهلة للقيام بذلك هي إدراج صف لأسماء المتغيرات في أعلى البيانات بملف إكسيل

ثم نسخة البيانات، بما في ذلك صف أسماء المتغيرات واستخدام لصق خاص Paste Special مع خيار معاملة الصف الأول كأسماء للمتغيرات Treat first tow as variable names وذلك لوضع كل البيانات في محرر بيانات خال.

طريقة ملف إكسيل ومحرر البيانات Data Editor هي طريقة سهلة وسريعة ولكن بالنسبة للبيانات الكبيرة، فإنه من الضروري أن تكون هذه البيانات محفوظة بواسطة برامج إحصائية أخرى مثل ملفات SAS أو SPSS أو SAS XPORT والتي يمكن استيراد بياناتها من خلال اختيارات قائمة ستاتا

File > Import > SAS XPORT

أو باستخدام الأمر **import sasxport**، كما أن تنسيقات البيانات الأخرى يمكن قراءتها باستخدام ملفات نصية وسيطة أو ترجمتها مباشرة باستخدام برامج خاصة.

بالإمكان شرح طريقة الملفات النصية باستخدام السلاسل الزمنية للمناخ، فالتردد الجنوبي إلنينيو (سوف يتم اختصارها ENSO) هي ظاهرة مناخية شبه دورية تحدث في المنطقة الاستوائية بالمحيط الهادئ، ولكنها أيضاً تؤثر على المناطق الأخرى، ويتم تفسير أجواء المنطقة الاستوائية بالمحيط الهادئ من خلال مؤشر ENSO المتعدد والذي يتضمن ستة متغيرات (الضغط على مستوى البحر، رياح السطح الجنوبية والإقليمية، سطح البحر ودرجة حرارة هواء سطح البحر، والسحاب) يتم دمجها في مؤشر واحد لـ ENSO، الملف النصي *MEI.raw* يحتوي على قيم شهرية لمؤشر ENSO للفترة من يناير 1950، وحتى ديسمبر 2011، وهذه القيم تم فصلها باستخدام مفتاح tab، وفاصلة لملف نصي تمت كتابته بواسطة أحد برامج الجداول الإلكترونية، الصف الأول من النص يحتوي على قائمة بأسماء المتغيرات وهي: *mei1* لقيمة المؤشر خلال شهر يناير، *mei2* لقيمة المؤشر خلال شهر فبراير وهكذا (في الواقع فإن قيمة المؤشر لشهر يناير تمثل ديسمبر – يناير، وقيمة المؤشر لشهر فبراير تمثل يناير – فبراير وهكذا) الصفوف الأولى من الملف النصي تظهر كما يلي:

```
year   mei1    mei2    mei3    mei4    mei5    mei6    mei7    mei8    mei9   mei10   mei11   mei12
1950   -1.022  -1.148  -1.287  -1.058  -1.423  -1.363  -1.342  -1.066  -.576  -.394   -1.154  -1.247
1951   -1.068  -1.194  -1.216  -.434   -.264   .482    .756    .864    .779   .752    .728    .467
1952   .406    .142    .096    .261    -.257   -.633   -.235   -.157   .362   .311    -.338   -.125
1953   .024    .388    .272    .712    .833    .242    .421    .252    .522   .092    .049    .313
```

يمكننا قراءة هذه البيانات باستخدام ستاتا عن طريق الأمر **insheet** مع خيارات لتحديد أن القيم تم فصلها باستخدام مفتاح tab وأن الصف الأول يحتوي على أسماء المتغيرات. وبعد قراءة بيانات الصفوف، يمكننا حفظ الملف بتنسيق ستاتا، وسوف نقوم بتسميته *MEI0.dta* وسوف نستخدم هذا الملف لاحقاً.

```
.insheet using c:\data\MEI.raw, tab name clear
.save c:\data\MEI0.dta, replace
.describe
Contains data from c:\data\MEI0.dta
  obs:            62
 vars:            13                          29 Apr 2013 17:01
 size:         3,100
--------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
--------------------------------------------------------------------------------
year            int    %8.0g
mei1            float  %9.0g
mei2            float  %9.0g
mei3            float  %9.0g
mei4            float  %9.0g
mei5            float  %9.0g
mei6            float  %9.0g
mei7            float  %9.0g
mei8            float  %9.0g
mei9            float  %9.0g
mei10           float  %9.0g
mei11           float  %9.0g
mei12           float  %9.0g
--------------------------------------------------------------------------------
Sorted by:
```

بإضافة الخيار **comma** بدلاً من **tab** و**insheet** يمكن قراءة ملف نصي يحتوي على قيم بينها فواصل، وهذا النوع من الملفات النصية هو الصيغة الأكثر شيوعاً لمخرجات الجداول الإليكترونية، ويمكن أيضاً قراءة الملفات النصية باستخدام قوائم برنامج ستاتا، ولمشاهدة الخيارات المتوافرة قم باختيار Data > Import.

حتى الآن الأمثلة التي تم شرحها افترضت أن قيم البيانات تم فصلها عن بعضها بفواصل أو مفتاح tab أو محددات معلومة أخرى (والتي يمكن أن يتم استخدامها مع الفواصل ومفتاح tab)، هناك إجراء آخر مختلف يُطلق عليه

تنسيق العمود الثابت fixed-column format حيث تكون البيانات غير مفصولة بعلامات معينة، ولكنها تقع في مواقع أعمدة محددة مسبقاً، ومثل هذا النوع من الملفات يمكن قراءته باستخدام الأمر **infix**، حيث يجب علينا أولاً تحديد كيفية قراءة الأعمدة.

فعلى سبيل المثال، لدينا بيانات مخزنة بملف نصي من نوع (ASCII) تحت اسم *nfresour.raw*

```
198624087641691000
198725247430001044
198825138637481086
198925358964371140
1990    8615731195
1991    7930001262
```

البيانات أعلاه تتعلق بإنتاج الموارد الطبيعية في نيوفوندلاند بكندا، المتغيرات الأربعة تقع في مواقع أعمدة ثابتة، فالأعمدة من (1 – 4) تمثل السنة (1986 .. 1991)، الأعمدة من 5 – 8 حجم إنتاج الغابات بالمتر المكعب بالآلاف (2408 ... قيمة مفقودة)، الأعمدة من 9 – 14 حجم إنتاج المناجم بالدولار بالآلاف (764,169 .. 793,000) أما الأعمدة من 15 – 18 فإنها مؤشر أسعار المستهلكين، ففي سنة 1986 كان المؤشر 1000 وفي سنة 1991 كانت قيمة المؤشر 1262، يجب ملاحظة أن الفراغات في تنسيق الأعمدة السابقة تعني وجود قيم مفقودة، كما أن البيانات لاتحتوي على أي فواصل عشرية. ولاستيراد بيانات *nfresour.raw* داخل برنامج ستاتا يجب علينا تحديد موقع عمود كل متغير كما يلي:

```
.infix year 1-4 wood 5-8 mines 9-14 CPI 15-18
using "C:\Data\nfresour.raw", clear
.list
```

	year	wood	mines	CPI
1.	1986	2408	764169	1000
2.	1987	2524	743000	1044
3.	1988	2513	863748	1086
4.	1989	2535	896437	1140
5.	1990	.	861573	1195
6.	1991	.	793000	1262

التنسيقات الأكثر تعقيداً من الأعمدة الثابتة، قد تتطلب قاموس بيانــات، وقواميس البيانات يمكن أن تكون واضحة وصــريحة وبهــا العديــد مــن الاختيارات، ولمعرفة المزيد من المعلومات عن هذه الخيارات، قم بطباعــة الأمر **help import**، ولمزيد من الأمثلة والشرح يمكنك الرجوع إلى دليــل مستخدم برنامج ستاتا؛ ويمكن لبرنامج ستاتا أيضاً استيراد بيانات من قواعد البيانات ODBC، وللحصول على معلومات أكثر عن ذلك، قم بطباعة الأمر **help odbc**.

ولكن ماذا إذا كنا نريد بيانات من برنامج ستاتا لاستخدامها في بــرامج أخرى؟ يمكن القيام بذلك عن طريــق الأمــر **export excel** والأمــر **export sasxport** كما أن قوائم ستاتا توفر العديد من الاختيارات وذلك من خلال

File > Export > Excel spreadsheet (*.xls; *.xlsx)
File > Export > SAS XPORT

الأوامر أعلاه، سوف تقوم بإنــشاء ملفــات إكسيل أو ملفــات SAS XPORT، كما يمكن أيضاً إنشاء ملفات نصية بتنسيقات مختلفة عن طريــق الأمر **outsheet** والأمر **outfile** (أو من خلال القائمــة Data > Export)، الخيار الآخر الأكثر ســرعة هو نســخ البيانات من نافذة محرر البيانــات Data Editor أو نافذة متصفح البيانات Data Browser ببرنامج ستاتا ولصقها مباشرة في برنامج الجداول الإليكترونية مثل برنامج إكسيل، وفي الغالب فإن أفضل خيار هو نقل البيانات مباشرة بين الملفات المخزنة ببرامج إحصائية أو قواعد بيانات متخصصة، وهناك بعض البرامج الأخرى التي قد تقوم بتحويل البيانات إلى صيغ مفهومة للبرامج الإحصائية، فمثلاً برنامج ستاتا ترانسفير Stata/Transfer تقوم بتحــويل البيانات إلى تنسيقــات مختلفة منها dBASE, Excel, FoxPro, Gauss, JMP, MATLAB, Minitab, OSIRIS, Paradox,R, S-Plus, SAS, SPSS, SYSTAT, Stata، حتى قواعد البيانات الكبيرة التــي تحتوي على كمية ضخمة من البيانــات يمكــن تحويلهــا بــســرعة مــع Stata/Transfer، وهــذا البرنامــج متوافــر مــن شركة ستاتا StataCorp (www.stata.com) أو مــن خــلال صــانع البرنامج وهو Circle Systems

(www.stattransfer.com)، وبرامج تحويل البيانات ضرورية عند العمل مع عدة برامج لتبادل البيانات مع الآخرين.

إحدى المميزات المهمة لبرنامج ستاتا، والتي يجب الإشارة إليها هي أن أي ملفات يتم حفظها ببرنامج ستاتا في أنظمة التشغيل المختلفة (سواء كانت Windows أو Mac أو Unix) فإنه بالإمكان فتحها ببرنامج ستاتا بدون الحاجة إلى تحويلها لتلائم نظام تشغيل معين، ولتحويل ملف بيانات تم حفظه بإصدار سابق من برنامج ستاتا إلى أحدث إصدارات ستاتا يجب استخدام الأمر **saveold** بدلاً من استخدام الأمر **save** أو استخدام قوائم ستاتا واختيار

File > Save As > Save as type > Stata 9/10 Data

دمج ملفين ستاتا أو أكثر : Combining Two or More Stata Files

بشكل عام، يمكننا دمج ملفات ستاتا بطريقتين: إرفاق **append** ملف البيانات الثاني، والذي يحتوي على المشاهدات الإضافية أو دمجه **merge** مع الملف الذي يحتوي على المتغيرات أو القيم الجديدة، فمثلاً الملف *lakewin1.dta* يحتوي على بيانات عن بداية ذوبان الجليد في أكبر بحيرة (بحيرة وينيبسوكي) بولاية هامبشير بالولايات المتحدة، وهذه البيانات تم تسجيلها بواسطة المواطنين القاطنين لمدة 121 سنة في فترة تمتد من 1887 إلى 2007.

```
. use c:\data\lakewin1.dta, clear
.describe

Contains data from c:\data\lakewin1.dta
  obs:          121                      Lake Winnipesaukee ice-out 1887-2007
 vars:            3                      2 Jul 2012 06:11
 size:          726
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
year            int    %ty                    Year
winedate        int    %tdCYmd                Lake Winnipesaukee Ice-Out
winout          int    %9.0g                  Lake Winnipesaukee Ice-Out day
-------------------------------------------------------------------------------
Sorted by:  year
```

```
.list in -4/1
```

	year	winedate	winout
118.	2004	2004Apr20	111
119.	2005	2005Apr20	110
120.	2006	2006Apr3	93
121.	2007	2007Apr23	113

من الجدول أعلاه، نجد أنه في سنة 2007 بدأ ذوبان الجليد في البحيرة في 23 أبريل، وهو اليوم 113 من السنة.

الملف *lakewin2.dta*: يحتوي على بيانات جديدة للفترة ما بين 2008 إلى 2012، وهو يحتوي على نفس المتغيرات بنفس التنسيق، لذا يمكننا دمج وتحديث الملف *lakewin2.dta* مع البيانات الجديدة الموجودة في الملف *lakewin1.dta*، باستخدام الأمر **append** كما يلي:

```
.use c:\data\lakewin2.dta
(Lake Winnipesaukee ice-out 2008-2012)
.describe

Contains data from c:\data\lakewin2.dta
  obs:           5                          Lake Winnipesaukee ice-out 2008-2012
 vars:           3                          2 Jul 2012 06:11
 size:          30
-------------------------------------------------------------------------------
              storage  display    value
variable name   type   format     label      variable label
-------------------------------------------------------------------------------
year            int    %ty                   Year
winedate        int    %tdCYmd               Lake Winnipesaukee Ice-Out
winout          int    %9.0g                 Lake Winnipesaukee Ice-Out day
-------------------------------------------------------------------------------
Sorted by:  year

.list
```

	year	winedate	winout
1.	2008	2008Apr23	114
2.	2009	2009Apr12	102
3.	2010	2010Mar24	83
4.	2011	2011Apr19	109
5.	2012	2012Mar23	83

```
.append using c:\data\lakewin1
.sort year
.label data "Lake Winnipesaukee ice out 1887-
 2012"
.save c:\data\lakewin3
.list in -7/1
```

	year	winedate	winout
120.	2006	2006Apr3	93
121.	2007	2007Apr23	113
122.	2008	2008Apr23	114
123.	2009	2009Apr12	102
124.	2010	2010Mar24	83
125.	2011	2011Apr19	109
126.	2012	2012Mar23	83

في هذا المثال البيانات في الملفين الاثنين لها نفس المتغيرات بالرغم من أنه ليس من الضروري للأمر **append** أن يحتوي الملفين على نفس المتغيرات، فإذا وجدت هناك متغيرات في ملف واحد ولم توجد في الملف الآخر، فإن البيانات المضاف إليها سوف تعتبر البيانات غير الموجودة قيماً مفقودة عند إجراء عملية الإرفاق.

الأمر *append* : يشبه إلى حد ما ورقة طويلة تحتوي على بيانات تم إلصاق ورقة أخرى في أسفلها تحتوي على بيانات جديدة، وبذلك يزيد طول الورقة. وهذا يعني إضافة مشاهدات جديدة للمتغيرات، أما الأمر دمج **merge** ففي أبسط أشكاله يشبه إضافة ورقة جديدة على يمين الورقة الجديدة، ويزيد عرض الورقة، وهذا يعني إضافة متغيرات جديدة للبيانات.

الملف *lakesun.dta* : يحتوي على بيانات ذوبان الجليد لثاني أكبر بحيرة – بحيرة سنابي – بولاية نيوهامبشير خلال الفترة من 1869 إلى 2012، وبالرغم من أن بيانات أكبر بحيرة (*lakewin3.dta*)، وبيانات ثاني أكبر بحيرة (*lakesun.dta*) تم الحصول عليهما من مصادر مختلفة فإنهما يمثلان سلسلة بيانات سنوية يمكن بسهولة دمجهما في ملف بيانات واحد، وذلك باستخدام الأمر **merge 1:1 *year***.

```
.use c:\data\lakesun.dta
.describe

Contains data from c:\data\lakesun.dta
  obs:           144                          Lake Sunapee ice-out 1869-2012
 vars:             3                          2 Jul 2012 06:11
 size:         1,152

              storage  display     value
variable name   type   format      label      variable label

year            int    %ty                    Year
sunedate        float  %tdCYmd                Date Lake Sunapee Ice-Out
sunout          int    %9.0g                  Lake Sunapee Ice-Out day

Sorted by:  year
```

```
.merge 1:1 year using c:\data\lakewin3.dta

    Result                           # of obs.

    not matched                            18
        from master                        18  (_merge==1)
        from using                          0  (_merge==2)

    matched                               126  (_merge==3)
```

كلا ملفي البيانات تم ترتيبهما حسب السنة، وإذا لم يكن هذا هو الوضع فيجب علينا أن نقوم بالترتيب حسب السنة **sort** ***year*** قبل إجراء عملية الدمج، نتائج الدمج توضح أن هناك 126 سنة في كلا الملفين (يجب ملاحظة أن البيانات الموجودة في ذاكرة ستاتا الآن هي بيانات الملف الرئيس *lakesun.dta*)، والبيانات المستخدمة في عملية الدمج كانت من الملف المصدر *lakewin3.dta*، وهناك 18 سنة إضافية (1869 إلى 1886) فقط موجودة في الملف *lakesun.dta*، لذلك فإن متغيرات بحيرة وينيبسوكي (أكبر بحيرة) سوف تحتوي على قيم مفقودة لهذه السنوات، ويقوم الأمر **merge** بإنشاء متغير اسمه *_merge* وهو يسجل ما إذا كانت المشاهدات تم الحصول عليها من الملف الرئيس فقط (*_merge=1*)، أو أن البيانات تم الحصول عليها من الملف المصدر (*_merge=2*)، أو أن البيانات من كلا الملفين (*_merge=3*)، وهناك خطوة مهمة جداً وهي معاينة القيم المدمجة ومراجعتها

بعناية بعد كل عملية دمج، وذلك للتأكيد على أن كل شيء تم حسب ما هو مخطط، وقبل القيام بعملية دمج أخرى يجب استبعاد **drop** أو إعادة تسمية **rename** المتغير الجديد *merge*_ وذلك كما يلي:

```
.drop _merge
.sort year
.list in 1/4
```

	year	sunedate	sunout	winedate	winout
1.	1869	1869May9	129	.	.
2.	1870	1870May9	129	.	.
3.	1871	1871Apr11	101	.	.
4.	1872	1872May2	123	.	.

```
.list in -4/1
```

	year	sunedate	sunout	winedate	winout
141.	2009	2009Apr11	101	2009Apr12	102
142.	2010	2010Apr3	93	2010Mar24	83
143.	2011	2011Apr21	111	2011Apr19	109
144.	2012	2012Mar22	82	2012Mar23	83

```
.label data "Sunapee & Winnipesaukee ice out
 1869-2012"
.save c:\data\lakesunwin.dta, replace
```

في هذا المثال قمنا باستخدام الأمر **merge** لإضافة متغيرات جديدة البيانات، وذلك بمقارنة المشاهدات مع السنة *year*، وبشكل افتراضي فإنه عند وجود نفس المتغيرات في ملفي بيانات، فإن البيانات الموجودة في الملف الرئيس تبقى كما هي، ويتم إهمال البيانات الموجودة في الملف المصدر. وهناك عدة خيارات يمكن استخدامها مع الأمر **merge** والتي يمكن أن تُعدّل الوضع الافتراضي عند دمج البيانات، فمثلاً الأمر أدناه، يسمح باستبدال أي قيمة مفقودة في الملف الرئيس بأي قيمة موجودة في الملف المصدر (ملف البيانات هنا اسمه *newdata.dta*).

```
.merge 1:1 year using newdata.dta, update
```

أو يمكن استخدام أمر آخر يقوم باستبدال أي قيمة في الملف الرئيس بأي قيمة غير مفقودة في الملف المصدر، وذلك إذا كانت قيم الملف المصدر مختلفة عن قيم الملف الرئيس.

```
.merge 1:1 year using newdata, update replace
```

كل هذه الأمثلة توضح كيفية دمج ملف مع ملف آخر (1 إلى 1)، ولكن هناك احتمالية لدمج ملف بعدة ملفات أخرى (1 إلى مجموعة) وذلك باستخدام الخيار **1:m**، أو دمج عدة ملفات بعدة ملفات أخرى بالخيار **m:m**. ولمزيد من المعلومات والأمثلة عن الدمج، قم بطباعة الأمر **help merge** وانظر دليل إدارة البيانات *Data Management Manual*، كما يمكن القيام بتعديل البيانات من خلال قوائم ستاتا باختيار Data > Combine datasets.

طي البيانات : Collapsing Data

بعد الحصول على البيانات وترتيبها، قد نكتشف لاحقاً أن هذه البيانات لاتلائم احتياجاتنا، أو أن هذه البيانات تحتوي على أخطاء، لحسن الحظ هناك العديد من الأوامر التي تسهل عملية إعادة بناء البيانات كما يجب، أسهل طريقة للقيام بذلك هي طي البيانات **collapse** حيث يتم تجميع البيانات في مجموعات يتم تعريفها بالمتوسط الحسابي أو الوسيط أو أي مقياس إحصائي آخر؛ ولتوضيح ذلك نقوم بالعودة إلى بيانات درجات الحرارة العالمية الشهرية في الفترة ما بين يناير 1880 إلى ديسمبر 2011 بالملف *global2.dta* والتي تم توضيحها بالرسم البياني رقم (1.2).

```
.use c:\data\global2.dta, clear
.describe
```

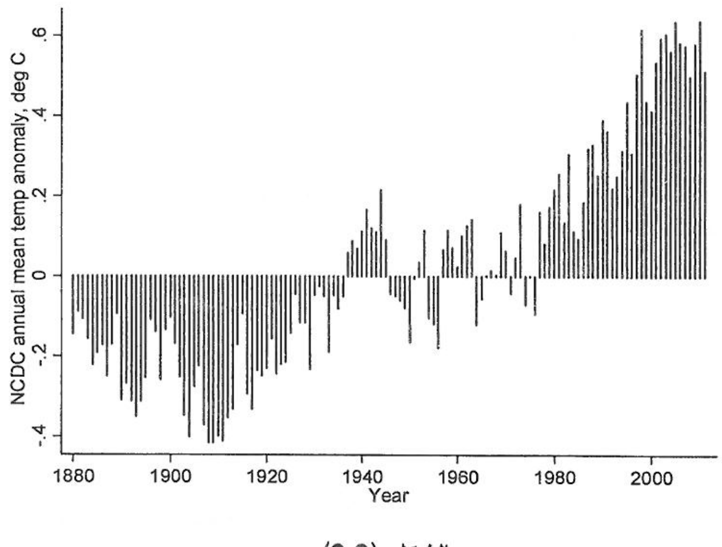
```
Contains data from c:\data\global2.dta
  obs:         1,584                          Global climate
 vars:             4                          4 Jul 2012 11:21
 size:        14,256
--------------------------------------------------------------------------------
              storage  display    value
variable name   type   format     label      variable label
--------------------------------------------------------------------------------
year            int    %8.0g                 Year
month           byte   %8.0g                 Month
edate           int    %tdmCY                elapsed date
temp            float  %9.0g                 NCDC global temp anomaly vs 1901-2000, C
--------------------------------------------------------------------------------
Sorted by:  edate
```

باستخدام الأمر **collapse** يمكننا بناء مجموعة بيانات بسيطة تتضمن متوسط انحراف درجات الحرارة لفترة 132 سنة بدلاً من 1584 شهراً منفصلاً.

```
.collapse (mean) temp, by(year)
.label variable temp "NCDC annual mean temp
 anomaly, deg C"
.save c:\data\global_yearly.dta, replace
.describe
```

```
Contains data from c:\data\global_yearly.dta
  obs:           132                          Global climate
 vars:             2                          30 Apr 2013 15:09
 size:           792
--------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
--------------------------------------------------------------------------------
year            int    %8.0g                  Year
temp            float  %9.0g                  NCDC annual mean temp anomaly, deg C
--------------------------------------------------------------------------------
Sorted by:  year
```

البيانات الجديدة يمكن توضيحها برسم بياني مضلع يشير إلى انحراف درجات الحرارة السنوية أعلى أو أقل من المتوسط خلال الفترة من 1901 إلى 2000.



الشكل (2.2)

الأمر **collapse**: يمكنه إنشاء متغيرات طبقاً للحسابات الإحصائية التالية:

mean	المتوسط الحسابي (وهو الخيار الافتراضي إذا لم يتم تحديد أي شيء آخر).
median	الوسيط
p1	المئين الأول
p2	المئين الثاني
sd	الانحراف المعياري
semean	الخطأ المعياري للمتوسط الحسابي((sd/sqrt(n))
sebinomial	الخطأ المعياري للمتوسط الحسابي، ذي الحدين ((sqrt(p(1-p)/n))
sepoisson	الخطأ المعياري للمتوسط الحسابي، بواسون ((sqrt(mean))
sum	المجاميع
rawsum	المجاميع مع تجاهل قيم استثنائية يتم تحديدها
count	عدد المشاهدات الموجودة فعلاً باستثناء القيم المفقودة
Max	أعلى قيمة
Min	أقل قيمة
Iqr	المدى الربيعي
First	أول القيم
Last	آخر القيم
Firstnm	قيمة أول مشاهدة موجودة (يتم استثناء القيم المفقودة).
Lastnum	قيمة آخر مشاهدة موجودة (يتم استثناء القيم المفقودة).

هناك عدد كبير من الحسابات الإحصائية يمكن إجراؤها باستخدام أمر أكثر مرونة وهو الأمر **statsby** وتتم كتابته قبل كتابة الحسابات الإحصائية المطلوبة. فعلى سبيل المثال، وباستخدام بيانات الملف *global2.dta* نقوم بإنشاء متغير جديد يسمى *decade* وهو يساوي 1880 للسنوات من 1880–1889، و1890 للسنوات 1890–1899 وهكذا، بعد ذلك نقوم بإنشاء ملف بيانات آخر يلخّص الإحصائيات المهمة لدرجات الحرارة خلال عقد من الزمن.

```
.use C:\data\global2.dta, clear
.gen decade = 10*int(year/10)
.statsby, by(decade) clear:summarize temp
```

```
(running summarize on estimation sample)

      command:  summarize temp
            N:  r(N)
        sum_w:  r(sum_w)
         mean:  r(mean)
          Var:  r(Var)
           sd:  r(sd)
          min:  r(min)
          max:  r(max)
          sum:  r(sum)
           by:  decade

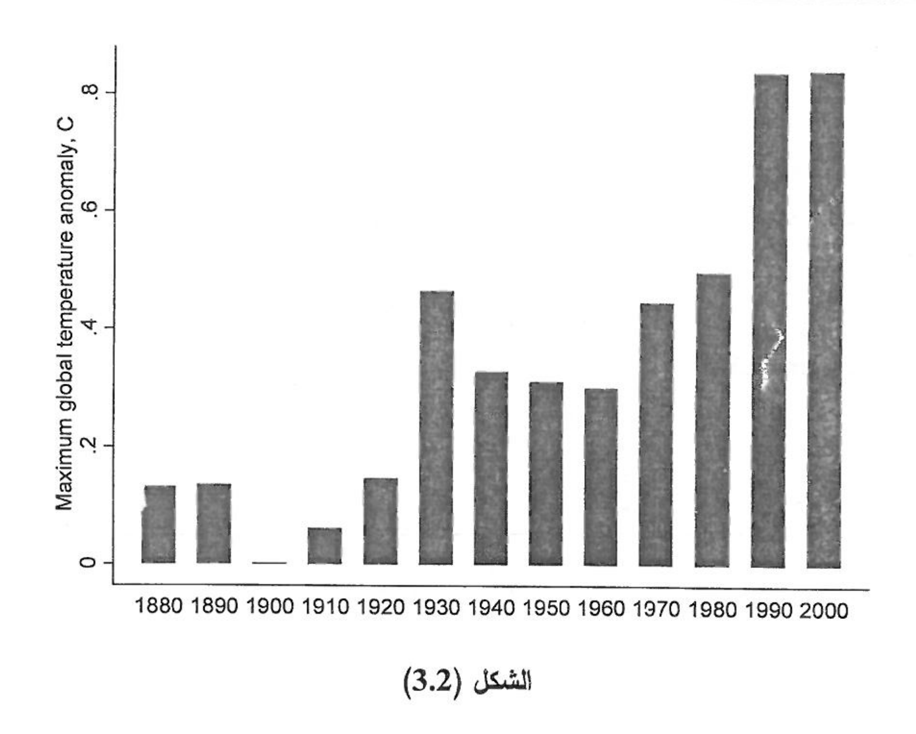
Statsby groups
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..............
```

البيانات الأخيرة تتضمن عدد المشاهدات، والمتوسط الحسابي، والتباين، وأعلى قيمة، وملخص بإحصائيات أخرى لكل عقد من العقود الزمنية في البيانات، الشكل (3.2) يعرض رسماً بيانياً بأعلى انحراف لدرجة الحرارة الشهرية لكل عقد (باستثناء عقد 2010 لأنه يتضمن سنتين فقط).

.describe

```
Contains data
  obs:          14                          statsby: summarize
 vars:           9
 size:         504
--------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
--------------------------------------------------------------------------
decade          float  %9.0g
N               float  %9.0g                  r(N)
sum_w           float  %9.0g                  r(sum_w)
mean            float  %9.0g                  r(mean)
Var             float  %9.0g                  r(Var)
sd              float  %9.0g                  r(sd)
min             float  %9.0g                  r(min)
max             float  %9.0g                  r(max)
sum             float  %9.0g                  r(sum)
--------------------------------------------------------------------------
Sorted by:
     Note:  dataset has changed since last saved
```

.graph bar *max* if *year*<2010, over(*decade*) ytitle("Maximum global temperature anomaly, C")



الشكل (3.2)

الأمر **statsby** يمكنه أيضاً إنشاء بيانات من نتائج معادلات الارتباط أو أي تحليلات أخرى، لمزيد من التفاصيل والأمثلة عن هذا الأمر، قم بطباعة **help statsby**، كما أن دليل مرجع إدارة البيانات *Data Management Reference Manual* يعرض معلومات عن هذا الأمر.

مستخدماً قوائم ستاتا قم بالاختيار التالي:

Statistics > Other > Collect statistics for a command across a by list

سوف فتح نافذة لهذا الأمر، وهناك أمر مفيد آخر وهو **contract** الذي يمكنه إنشاء بيانات بطريقه مشابهة للجدول التكراري لعدد من المتغيرات التي يتم تحديدها (انظر **help contract**).

إعادة تشكيل البيانات : Reshaping Data

هناك عدة طرق مختلفة يمكن القيام بها لإعادة تشكيل البيانات باستخدام الأمر **reshape**، هذا الأمر يُرتب البيانات بطرق بسيطة بين ترتيب طولي

وترتيب عرضي. ففي جزء سابق من هذا الفصل، قمنا بإنشاء ملف يحتوي على بيانات عن مؤشر ENSO المتعدد (*MEI0.dta*)، حيث إن البيانات في تنسيق عريض: السنوات تظهر في صفوف، ولكن كل شهر في عمود منفصل، حيث إن العمود *mei1* يمثل قيمة المؤشر في شهر يناير، *mei2* يمثل قيمة المؤشر في شهر فبراير وهكذا.

.use C:\data*MEI0.dta*, clear
.describe

```
Contains data from c:\data\MEI0.dta
  obs:            62
 vars:            13                          29 Apr 2013 17:51
 size:         3,100

              storage  display     value
variable name   type   format      label      variable label

year            int    %8.0g
mei1            float  %9.0g
mei2            float  %9.0g
mei3            float  %9.0g
mei4            float  %9.0g
mei5            float  %9.0g
mei6            float  %9.0g
mei7            float  %9.0g
mei8            float  %9.0g
mei9            float  %9.0g
mei10           float  %9.0g
mei11           float  %9.0g
mei12           float  %9.0g

Sorted by:
```

.list year-mei7 in 1/5

	year	mei1	mei2	mei3	mei4	mei5	mei6	mei7
1.	1950	-1.022	-1.148	-1.287	-1.058	-1.423	-1.363	-1.342
2.	1951	-1.068	-1.194	-1.216	-.434	-.264	.482	.756
3.	1952	.406	.142	.096	.261	-.257	-.633	-.235
4.	1953	.024	.388	.272	.712	.833	.242	.421
5.	1954	-.051	-.019	.169	-.504	-1.397	-1.578	-1 382

يمكننا إعادة تشكيل هذا التنسيق العريض للبيانات وجعلها سلسلة زمنية في تنسيق طولي، فالأمر أدناه يقوم بإنشاء متغير جديد باسم *mei*، وكل صف في البيانات الجديدة ذات التنسيق الطويل سوف يحتوي على محدد لكل مشاهدة i(*year*) ومحدد لكل مشاهدة فرعية j(*month*) .

```
.reshape long mei, i(year) j(month)
(note: j = 1 2 3 4 5 6 7 8 9 10 11 12)

Data                                   wide   ->   long
-----------------------------------------------------------
Number of obs.                           62   ->    744
Number of variables                      13   ->      3
j variable (12 values)                        ->   month
xij variables:
                      mei1 mei2 ... mei12     ->   mei
-----------------------------------------------------------

.compress
.sort year month
.label variable mei "Multivariate ENSO Index"
.save C:\data\mei1.dta, replace
.describe
Contains data from c:\data\mei1.dta
  obs:           744
 vars:             3                          30 Apr 2013 19:44
 size:         5,208
----------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
----------------------------------------------------------------------
year            int    %8.0g
month           byte   %9.0g
mei             float  %9.0g                  Multivariate ENSO Index
----------------------------------------------------------------------
Sorted by:  year  month

.list in 1/5

     +------------------------+
     | year   month      mei  |
     |------------------------|
  1. | 1950       1   -1.022  |
  2. | 1950       2   -1.148  |
  3. | 1950       3   -1.287  |
  4. | 1950       4   -1.058  |
  5. | 1950       5   -1.423  |
     +------------------------+
```

أصبح لدينا الآن سلسلة زمنية لمؤشر ENSO المتعدد على شكل سنوي/ شهري مشابه للسلسلة الزمنية الخاصة بدرجات الحرارة العالمية التي تم حفظها في ملف *global2.dta* سابقاً في هذا الفصل، مع وجود بيانات كلا الملفين مرتبة حسب السنة والشهر، يمكننا دمج الملفين في ملف واحد كما يلي:

```
.use c:\data\global2.dta, clear
.merge 1:1 year month using C:\data\mei1.dta
```

```
Result                           # of obs.
-----------------------------------------
not matched                           840
    from master                       840  (_merge==1)
    from using                          0  (_merge==2)

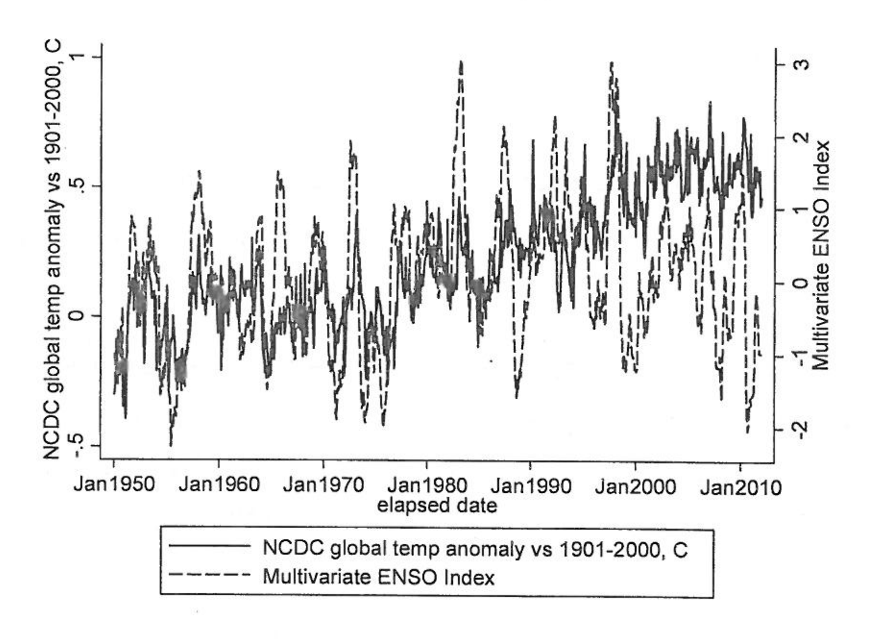
matched                               744  (_merge==3)
-----------------------------------------
```

بيانات درجات الحرارة في الملف *global2.dta* تغطي كل شهر للفترة ما بين يناير 1880 وحتى ديسمبر 2011، بينما البيانات بالملف *mei1.dta* تغطي فقط الفترة من يناير 1950 إلى ديسمبر 2011، وبالتالي لدينا 70×12=840 شهرًا موجودة فقط في الملف الرئيس، وهي لا تتطابق مع باقي البيانات، وهي عبارة عن 12×62=744 شهرًا موجودة في كلا الملفين وهي متطابقة.

بعد حفظ بيانات الملف الجديد باسم *global3.dta*، يمكننا أن نرسم شكلاً بيانياً للمدة الزمنية يوضح درجات الحرارة مع المتغير *mei* خلال السنوات من 1950 وحتى 2011، هذان المتغيران لهما قياسات مختلفة، فالمتغير *mei* سوف يكون على الجانب الأيمن لمحور *y* وهو يشير إلى المحور العمودي 2، الأمر **graph** أدناه يقوم بإنشاء رسم بياني يتضمن منحنيين، منحنى لدرجة الحرارة والآخر للمتغير *mei* وهي المدة. وتم رسم المنحنى الثاني بخط متقطع، كما أن الأمر يحدد أيضًا مربع شرح الرسم البياني بصفين بدلاً من الوضع الافتراضي وهو عمودان. النظرة الأولى للرسم البياني تشير إلى أن درجات الحرارة العالمية ومؤشر ENSO في الغالب يتغيران معاً من سنة

لأخرى، ولكن مؤشر ENSO كان أكثر انخفاضاً من درجات الحرارة خلال العقود الأخيرة. الفصل (12)، يتعامل مع مثل هذا النوع من نماذج السلاسل الزمنية بطريقة أكثر دقة.

```
.sort year month
.drop _merge
.compress
.save C:\data\global3.dta, replace
.graph twoway line temp edate ||
 line mei edate, yaxis(2) lpattern(dash) ||
 if  year>1949, legend(row(2))
```



الشكل (4.2)

الأمر reshapre: يقوم بنفس العمل عند استخدامه بطريقة عكسية لتغيير البيانات من التنسيق الطولي إلى التنسيق العرضي، حيث يمكننا تغيير السلسلة الزمنية للشهر والسنة الخاصة بدرجات الحرارة ومؤشر ENSO لتكون في وضع عرضي، بحيث يكون كل صف يمثل سنة، وكل عمود يمثل شهرًا وذلك باستخدام الأوامر التالية:

```
.drop edate
.reshape wide mei temp, i(year) j(month)
```

استخدام الأوزان : Using Weights

يمكن لبرنامج ستاتا قبول أربعة أنواع من الأوزان:

aweight الأوزان التحليلية والتي تُستخدم في ارتباط الربيعات الأقل وزناً والتحليلات المشابهة الأخرى.

fweight الأوزان التكرارية وتقوم بعد الأرقام المتكررة في المشاهدات، والأوزان التكرارية يجب أن تكون رقماً صحيحاً.

iweight الأوزان المهمة ويتم تعريفها حسب رغبة المستخدم.

pweight الأوزان الاحتمالية أو أوزان المعاينة وهي الجزء العكسي من احتمال أن مشاهدة ما تم إدراجها بسبب استراتيجية المعاينة.

كل الأوزان ليست معروفة لكل أنواع التحليلات، فلا يمكننا مثلاً استخدام **pweight** مع الأمر **tabulate**؛ الاستخدام الفعال للأوزان يتطلب فهماً كاملاً لما هو مطلوب من هذه الأوزان القيام به في أي تحليل.

للأوزان تطبيقات إحصائية عديدة منها طرق تعويض الأشياء غير المتناسبة، وتصميمات المعاينة المعقدة، وهي ميزة عامة للدراسات الاستقصائية، والوزن **pweight** يوفر طريقة لتعديل التحيّز في العينات باستخدام الأوزان الاحتمالية والتي تساوي 1/(احتمالية الاختيار)، تحليل بيانات الدراسات الاستقصائية باستخدام الأوزان الاحتمالية يعتبر نقطة قوة لبرنامج ستاتا، والتي سيتم عرضها في الفصل 4.

في بعض الأمثلة، عملية الوزن تشبه إلى حد ما تجميع البيانات بطريقة تجعل المتغيرات تلخص إحصائيات العديد من المشاهدات الفردية، فمثلاً بيانات الملف *Nations2.dta* تحتوي على مؤشرات التطور البشري للأمم المتحدة، والتي توضح ظروف المعيشة في 194 دولة.

```
.use C:\data\nations2.dta, clear
.describe
```

```
Contains data from c:\data\nations2.dta
  obs:           194                          UN Human Development Indicators
 vars:            13                          2 Jul 2012 06:11
 size:        12,804

              storage  display     value
variable name   type   format      label      variable label

country         str21  %21s                   Country
region          byte   %8.0g       region     Region
gdp             float  %9.0g                  Gross domestic product per cap
                                                2005$, 2006/2009
school          float  %9.0g                  Mean years schooling (adults)
                                                2005/2010
adfert          float  %8.0g                  Adolescent fertility: births/1000
                                                fem 15-19, 2010
chldmort        float  %9.0g                  Prob dying before age 5/1000 live
                                                births 2005/2009
life            float  %9.0g                  Life expectancy at birth 2005/2010
pop             float  %9.0g                  Population 2005/2010
urban           float  %9.0g                  Percent population urban 2005/2010
femlab          float  %9.0g                  Female/male ratio in labor force
                                                2005/2009
literacy        float  %9.0g                  Adult literacy rate 2005/2009
co2             float  %9.0g                  Tons of CO2 emitted per cap
                                                2005/2006
gini            float  %9.0g                  Gini coef income inequality
                                                2005/2009

Sorted by:  region  country
```

"متوسط" "Mean" العمر المتوقع هو 68.7 سنة:

```
.summarize life
```

Variable	Obs	Mean	Std. Dev.	Min	Max
life	194	68.7293	10.0554	45.85	82.76666

المتوسط أعلاه يمثل متوسط العمر المتوقع للسكان في 194 دولة بالعينة وليس لكل السكان 7 مليارات القاطنين في هذه الدول، حيث إن وزن متوسط العمر لأصغر دولة (توفانلو هي جزيرة صغيرة في المحيط الهادئ بعدد سكان يبلغ حوالي 10000 نسمة) هو نفسه متوسط العمر المتوقع لأكبر دولة (الصين وعدد سكانها حوالي 1.3 مليار نسمة) ولكن باستخدام عدد السكان كوزن تكراري يمكننا الحصول على توقعات أكثر دقة عن متوسط العمر المتوقع لعدد السكان 7 مليارات ككل.

```
.summarize life [fweight=pop]
```

Variable	Obs	Mean	Std. Dev.	Min	Max
life	6.669e+09	68.93644	8.095538	45.85	82.76666

الأوزان الاحتمالية (pweight) سوف تتم مناقشتها بتفصيل أكثر في الفصل 4، والأوزان التحليلية (aweight) مفيدة مع إنشاء الرسومات البيانية. وسيتم شرحها في الفصل 3، أما المربعات الصغرى الموزونة فسوف يتم شرحها في الفصلين 7 و8 مع بعض المواضيع الأخرى؛ أهمية الأوزان ليس لها تعريف محدد، ولكن يمكن تطبيقها في برامج مكتوبة خصيصاً لأغراض محددة.

إنشاء بيانات عشوائية وعينات عشوائية :

Creating Random Data and Random Samples

دالة الأرقام شبه العشوائية **runiform()** تأتي في مقدمة الدوال التي تعطي لبرنامج ستاتا القدرة على إنشاء بيانات عشوائية أو عينات عشوائية، دليل المستخدم *Base Reference Manual* يعرض شرحاً تقنياً لكيفية عمل دالة الأرقام شبه العشوائية، إذا كانت هناك بيانات في الذاكرة الحالية لبرنامج ستاتا، فإن أمرًا مثل الأمر أدناه يقوم بإنشاء متغير جديد باسم *random* بطريقة عشوائية بقيمة 16 رقم يفصل بينها [0,1).

```
.generate randnum=runiform()
```

كما يمكننا إنشاء بيانات عشوائية من العدم، وللقيام بذلك سوف نقوم أولاً بمسح البيانات الموجودة بالذاكرة الحالية لبرنامج ستاتا (إذا كانت هذه البيانات مهمة فيجب حفظها أولاً) ثم نقوم بتحديد عدد المشاهدات المرغوب بها في البيانات الجديدة، ثم تحديد تكرار الأرقام العشوائية مما يجعل عملية إنتاج نفس البيانات لاحقاً عملية سهلة، وأخيراً نقوم بإنشاء المتغير العشوائي، الأمر أدناه يقوم بإنشاء بيانات تحتوي على 10 مشاهدات ومتغير واحد يسمى *randnum*

```
.clear
.set obs 10
.set seed 12345
.generate randum = runiform()
.list
```

	randum
1.	.309106
2.	.6852276
3.	.1277815
4.	.5617244
5.	.3134516
6.	.5047374
7.	.7232868
8.	.4176817
9.	.6768828
10.	.3657581

بالجمع بين دوال الجبر والإحصاء الخاصة ببرنامج ستاتا **runiform()** يمكن إجراء قيم لعينة تم جمعها من توزيعات نظرية متنوعة، فإذا كنا نريد متغيرًا جديدًا لنفرض أن اسمه *newvar* لعينة تم جمعها من توزيع منتظم حتى [0,428) بدلاً من المعتاد وهو [0,1) فإننا نقوم بطباعة الأمر.

```
.generate newvar = 428 * runiform()
```

الأمر سوف يظل ينتج قيماً حتى 16 رقماً، ولكن قد نريد أعدادًا صحيحة فقط من 1 إلى 428، الدالة **ceil()** توفر طريقة سهلة للقيام بذلك:

```
.generate newvar=ceil(428*runiform())
```

ولإنشاء عينة عشوائية من 1000 مشاهدة وجدولها التكراري يتكون من 6 خانات نقوم بطباعة الأوامر التالية:

```
.clear
.set obs 1000
.generate roll = ceil(6*runiform())
.tabulate roll
```

roll	Freq.	Percent	Cum.
1	172	17.20	17.20
2	166	16.60	33.80
3	149	14.90	48.70
4	170	17.00	65.70
5	168	16.80	82.50
6	175	17.50	100.00
Total	1,000	100.00	

نظرياً نحن نتوقع أن نجد أن رقم 1 تكرر بنسبة 16.67% وأن رقم 2 بنسبة 16.67% وهكذا، ولكن في أي عينة مثل تلك التي لدينا الآن والتي بها 1000 مشاهدة، فإن نسبة كل مشاهدة سوف تتباين بشكل عشوائي حول القيم المتوقعة.

ولإنشاء عينة مجموعة تتكون من 1000 مشاهدة وجدول تكراري يتكون من زوجين من رقم 6 يكون الأمر.

```
.generate dice = ceil(6*runiform()) +
 ceil(6*runiform(()
.tabulate dice
```

dice	Freq.	Percent	Cum.
2	26	2.60	2.60
3	57	5.70	8.30
4	68	6.80	15.10
5	107	10.70	25.80
6	141	14.10	39.90
7	179	17.90	57.80
8	121	12.10	69.90
9	108	10.80	80.70
10	108	10.80	91.50
11	61	6.10	97.60
12	24	2.40	100.00
Total	1,000	100.00	

يمكننا أيضاً استخدام n_ لتكوين عينة صناعية، الأوامر التالية تقوم بإنشاء بيانات تتضمن 5000 مشاهدة ومتغير واحد باسم *index* يحتوي على قيم من 1 إلى 5000.

```
.set obs 5000
.generate index = _n
.summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
roll	1000	3.521	1.73279	1	6
dice	1000	6.953	2.429546	2	12
index	5000	2500.5	1443.52	1	5000

لإنشاء متغيرات عشوائية من توزيع طبيعي (توزيع جاوس) نستخدم دالة **rnormal()**، الأمر التالي يقوم بإنشاء بيانات تتضمن 2000 مشاهدة ومتغيرين اثنين هما: المتغير z من المجتمع N(0,1) والمتغير u من N(500,75)

```
.clear
.set obs 2000
.generate z=rnormal()
.generate u=rnormal(500,75)
```

المتوسط الفعلي للعينة والانحراف المعياري يختلف قليلاً عن القيم النظرية.

```
.summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
z	2000	-.0016761	1.013879	-3.645242	3.620961
u	2000	498.4816	73.5583	248.0803	739.0969

إذا كان المتغير z يتبع التوزيع الطبيعي، فإن $v=e^z$ تتبع التوزيع اللوغاريتمي الطبيعي، ولإنشاء متغير لوغاريتمي طبيعي v بناءً على توزيع طبيعي معياري نقوم بالتالي:

```
.generate v=exp(rnormal())
```

استخدام اللوغاريتمات يجعل قيم المتغير أكثر تناسقاً.

لإنشاء قيم للمتغير w والتي تم الحصول عليها عشوائياً من توزيع أسّي بمتوسط وانحراف معياري يساوي $\mu = \sigma = 3$

```
.generate w=-3 * ln(runiform())
```

بالنسبة للمتوسطات والانحرافات المعيارية الأخرى، سوف يتم الاستعاضة عنها بثلاثة متغيرات أخرى.

x5 يتبع توزيع كاي تربيع مع درجة حرية تساوي 5

```
.generate x5 = rchi2(5)
```

y يتبع توزيع ذا الحدين وتم إعطاؤه 10 محاولات مع نسبة احتمال نجاح 0.2

```
.generate y = rbinomial(10,.2)
```

t45 يتبع توزيع *t* مع درجة حرية تساوي 45

```
.generate t45 = rt(45)
```

قم بطباعة الأمر **help random** للحصول على قائمة كاملة بالدوال الأخرى المتوافرة لإنشاء متغيرات عشوائية من توزيع بيتا، وتوزيع جاما، والتوزيع فوق الهندسي، والتوزيع ذي الحدين، وتوزيع بواسون.

الأمر drawnorm: يوفر طريقة أخرى لإنشاء متغيرات طبيعية متعددة وبه خيارات لتحديد الارتباط بين تلك المتغيرات، باستخدام الأمر **drawnorm** سوف يتم إنشاء بيانات تحتوي على 5000 مشاهدة لمتغير واحد من N(0,1) كما يلي:

```
.clear
.drawnorm z, n(5000)
.summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
z	5000	.0175441	.9919173	-4.112056	3.37486

الآن سوف نقوم بإنشاء ثلاثة متغيرات أخرى: المتغير *x1* من المجتمع N(0,1)، المتغير *x2* من N(100,15)، المتغير *x3* من N(500,75)، بالإضافة إلى ذلك سوف نعطي لهذه المتغيرات ارتباطات المجتمع التالية:

	x1	x2	x3
x1	1	0.4	−0.8
x2	0.4	1	0
x3	−0.8	0	1

عملية إنشاء مثل هذه البيانات تتطلب أولاً تحديد مصفوفة الارتبـــاط *C* واستخدام *C* مع الأمر **drawnorm**.

```
.mat C = (1, .4, -.8 \ .4, 1, 0 \ -.8, 0, 1)
.drawnorm    x1    x2    x3,    means(0,100,500)
 sds(1,15,75) corr(C)
.summarize x1-x3
```

Variable	Obs	Mean	Std. Dev.	Min	Max
x1	5000	-.0268606	.9969571	-3.635491	3.657701
x2	5000	99.77643	14.90536	42.78308	157.9469
x3	5000	500.2264	76.41673	227.5245	731.9073

```
.correlate x1-x3
(obs=5000)
```

	x1	x2	x3
x1	1.0000		
x2	0.3787	1.0000	
x3	-0.8024	0.0181	1.0000

قارن بين ارتباط المتغيرات ومتوسطاتها بالعينة مـــع القـــيم النظريـــة المعطاة سابقاً. البيانات العشوائية التي تم إنشاؤها بهذا النمط يمكن اعتبارها كعينة تم سحبها من مجتمعات نظرية ولا يجب أن نتوقع أن هذه العينات لها نفس سمات المجتمعات بالضبط (في هذا المثال سمات العينة كانت: متوسط المتغير *x3* يساوي 500، وارتباط المتغيرين *x1-x2* يساوي 0.4، وارتبـــاط المتغيرين *x1-x3* يساوي 0.80− وهكذا)، العينات الـــصناعية المترابطـــة أو غير المترابطة يمكن إنشاؤها باستخدام قوائم ستاتا، ونوافذ الحوار التي يمكن الحصول عليها كما يلي:

Statistics > Other > Draw a sample from a normal distribution

أو من خلال:

Statistics > Other > Create a dataset with specified correlation structure

الأمر sample: يقوم باستخدام بسيط لدالة **runiform** لإنشاء عينات عشوائية لبيانات من ذاكرة ستاتا. فعلى سبيل المثال، لإلغاء كل بيانات العينة مع الإبقاء على نسبة 10% فقط من البيانات الأصلية قم بطباعة الأمر:

```
.sample 10
```

وعند إضافة المحدد **in** أو المحدد **if**، فإن الأمر **sample** يقوم بالإبقاء فقط على المشاهدات التي تنطبق عليها المعايير المطلوبة:

```
.sample 10 if age< 26
```

الأمر أعلاه سوف يُبقي على نسبة 10% من المشاهدات كحد أدنى والتي فيها متغير *age* أكبر من أو يساوي 26، وإذا كانت المشاهدات التي قيمتها أكبر من أو تساوي 26، فسوف يتم الإبقاء عليها حتى وإن جاوزت 10%.

يمكننا اختيار عينات عشوائية من حجم معين،لإلغاء كل المشاهدات التي تم اختيارها عشوائياً من بيانات في ذاكرة ستاتا نقوم بطباعة الأمر.

```
.sample 90, count
```

الجزء الخاص بمحاكاة مونت كارلو في الفصل (14) يشرح بشكل تفصيلي أكثر عملية إنشاء متغيرات عشوائية.

كتابة برامج لإدارة البيانات :

Writing Programs for Data Management

إدارة البيانات في المشروعات الكبيرة تتضمن في العادة مهمات تكرارية أو قابلة لحدوث الأخطاء فيها، ويمكن لبرنامج ستاتا التعامل معها من خلال كتابة برامج خاصة، البرمجة المتقدمة يمكن أن تكون أكثر تعقيداً، ولكننا

يمكن أن نبدأ بكتابة برامج بسيطة تحتوي فقط على تسلسل لأوامر ستاتا يتم كتابتها وحفظها في ملف نصي، والملفات النصية يمكن إنشاؤها باستخدام أي برنامج محرر نصي يُفضله المستخدم أو برنامج المفكرة والذي يوفر عدة أنواع من الملفات النصية التي تحتوي على خيارات مختلفة عند الحفظ File <- Save As، إحدى الطرق السهلة لإنشاء ملفات نصية يتم باستخدام نافذة Do-file Editor ببرنامج ستاتا والتي يمكن الوصول إليها عن طريق اختيار Window > Do-file Editor أو بالضغط على أيقونة ، أو من خلال طباعة الأمر **doedit** أو **doedit *filename*** إذا كان هناك أي ملف *filename* نصي سبق تخزينه، الأوامر في نافذة المعاينة Review Window يمكن تحديدها وإرسالها مباشرة إلى نافذة Do-File Editor (انقر زر الفارة الأيمن للحصول على قائمة الاختيارات)، الأوامر يمكن نسخها ولصقها في نافذة Do-File Editor من أي مصادر أخرى مثل ملفات التسجيل أو نافذة النتائج.

في العديد من إجزاء هذا الفصل بدأنا ببناء بيانات درجات الحرارة العالمية بدءًا من درجة الحرارة، ثم إعادة تشكيل ودمج مؤشر ENSO وأخيراً تمثيل المؤشر ودرجات الحرارة في رسم بياني في الشكل (4.2). الأوامر التي قامت بتنفيذ كل هذه الخطوات يمكن تجميعها في ملف نصي واحد كما سيتم شرحه الآن، يجب ملاحظة أن استخدام /// للإشارة بأن الأمر **graph twoway** يستمر لأكثر من سطر واحد. في النهاية سوف يتم حفظ الرسم البياني في الشكل (4.2) بالامتداد (.emf)

```
insheet using c:\data\glbal.csv, comma clear
label data "Grlbal climate"
label variable year "Year"
label variable month "Month"
label variable temp "NCDC global temp anomaly
vs 1901-2000, C"
generate edate = mdy(month, 15, year)
label variable edate "elapsed date"
format edate %tdmCY
sort year month
order year month edate
save c:\dataglobal2.dta, replace
```

```
use c:\data\MEI0.dta, clear
reshape long mei, i(year) j(month)
sort year month
label variable mei "Multivariate ENSO Index"
save c:\data\mei1.dta, replace
use c:\data\global2.dta, clear
merge 1:1 year month using c:\data\mei1.dta
sort year month
drop _merge
compress
save c:\data\global3.dta, replace
graph twoway line temp edate///
 || line mei edate, yaxis(2) lpattern(dash)  ///
 || if year>1949, legend(row(2))
graph  save  Graph  "C:\graphs\fig02_04.gph",
replace
graph  export  "C:\graphs\fig02_04.emf",  as(emf)
replace
```

هذا الملف سوف يتم إعداده بتحديد الأوامر في نافذة المعاينة ثم اضغط على الزر الأيمن للفارة واختر Send to Do-file Editor ثم قم بحفظ ملف do-file باسم جديد مثل *global.do*، من القائمة أو بكتابة الأمر:

.do *global*

مثل هذا النوع من البرامج في العادة يتم حفظه بامتداد ".do" وهناك برامج أكثر دقة (يتم إعدادها بواسطة ملفات do-files أو تلقائياً بواسطة ملفات ado-files) يمكن حفظها في ذاكرة ستاتا وتشغيل هذا الملف على برامج أخرى وإنشاء أوامر جديدة لبرنامج ستاتا وفتح ملف وورد للأخطاء المحتملة.

برنامج ستاتا يقوم بتفسير نهاية سطر الأمر كنهاية للأمر نفسه، هذا معقول على الشاشة عندما يكون طول الأمر على عرض الشاشة ولكن قد يكون الأمر طويلاً جداً بحيث لن يعمل عند طباعته في ملف نصي، لذلك فإن وجود ثلاث شرطات مائلة (///) في نهاية سطر الأمر يشير إلى أن الأمر مازال مستمراً في السطر التالي، ولن يتم تنفيذ الأمر في حالة الوصول إلى

نهاية سطر الأمر مع ملاحظة أن وجود (///) لا يعني الوصول إلى نهاية الأمر.

هناك طريقة أخرى للتعامل مع الأوامر الطويلة في ملفات do-files وذلك باستخدام الأمر **#delimit;** حيث يحتوي على فاصلة منقوطة (;) في نهاية الأمر. في المثال أدناه، نقوم باستخدام الفاصلة المنقوطة، فبعد طباعة أمر طويل لا ينتهي حتى ظهور الفاصلة المنقوطة، وبعد ذلك نعدل المحددات إلى قيمتها الاعتيادية باستخدام المحدد (cr)

```
#delimit ;
Graph twoway line temp edate
  || line mei edate, yaxis(2) lpattern(dash)
  || if year>1949, legend(row(2)) ;
#delimit cr
```

في كل مرة يصل فيها برنامج ستاتا لنافذة النتائج يتوقف حتى نقوم بالضغط على مسافة في لوحة المفاتيح أو أي مفتاح آخر أو اضغط على أيقونة فبدلاً من التوقف يمكن الطلب من برنامج ستاتا الاستمرار في العمل حتى إظهار النتائج بالكامل وذلك بطباعة الأمر التالي:

```
.set more off
```

هذا الأمر يعتبر مناسباً إذا كان البرنامج يعرض مخرجات بحجم كبير على الشاشة، ونحن لا نحتاج إلى رؤية هذه المخرجات أو نحن نقوم بكتابة ملف تسجيل وسوف تتم مراجعة هذا الملف لاحقاً. وللرجوع إلى الوضع السابق بحيث يجب علينا النقر على أي مفتاح لإظهار بقية النتائج نطبع الأمر.

```
.set more on
```

الفصل الثالث

الرسومات البيانية

Graphs

تظهر الرسومات البيانية في كل فصل بهذا الكتاب، وهذا مؤشر على أهميتها وتداخلها مع التحليلات الأخرى ببرنامج ستاتا. الرسومات البيانية التحليلية تعتبردائماً إحدى نقاط القوة ببرنامج ستاتا، والسبب أن هناك العديد من الخيارات التي يتيحها برنامج ستاتا، ولا توجد في البرامج الإحصائية الأخرى. كما أن هذه الرسومات البيانية جذابة وقابلة للنشر وعملية إدراجها سهلة، وتتم من خلال استخدام الأوامر أو استخدام قوائم ستاتا باختيار Graphics، وبالنسبة المستخدمين الذين يتصورون الحصول على رسومات بيانية دقيقة ورائعة سوف يجدون أن ما يتصورونه واقعاً مع وجود العديد من الأدوات والخيارات المثيرة والمشروحة بتفصيل أكثر في دليل المستخدم *Graphics Reference Manual* وهناك العديد من الأمثلة في الدليل المرئي للرسومات البيانية ببرنامج ستاتا للمؤلف *A Visual Guide to* (Mitchell 2012) *Stata Graphics*.

في هذا الفصل، سوف يتم شرح عدة أنواع من الرسومات البيانية الرئيسة ببرنامج ستاتا وسيكون الشرح بأمثلة عملية بدلاً من الشرح النظري فقط (انظر دليل المستخدم *Graphics Reference Manual* أو استخدم ملفات المساعدة **help** للحصول على شرح نظري أكثر). بعض من أمثلة هذا الفصل سهلة جداً، وتستخدم خيارات قليلة، وأحياناً لاتستخدم أي خيارات ولكن عادة الأمثلة البسيطة يتبعها أمثلة أكثر تعقيداً مع بعض الخيارات التوضيحية وشرح عن كيفية إعداد الرسومات البيانية الجاهزة للنشر، وبالرغم من أن هذا الفصل يمكن المرور عليه سريعاً بقراءة كل مثال، ولكنه

يعتبر معرضاً للصور والأفكار التي يمكن استخدامها في إعداد الرسومات البيانية.

المجموعة الكاملة للخيارات المتعلقة بالرسومات البيانية هائلة جداً ولا يمكن لهذا الكتاب أن يغطيها بالشرح ولكن أمثلة هذا الفصل توضح مجموعة قليلة من هذه الخيارات، وفي فصول لاحقة سوف يتم الشرح بصورة أكثر تفصيلاً، وتوفر قائمة ستاتا Graphics الخيارات التي يمكن النقر عليها للوصول إلى قوائم الرسومات البيانية مع نوافذ حوار توجد بكل نافذة منها أيقونة Submit، هذه القائمة تعتبر طريقة مفيدة لتعرف ما هو متوافر من خيارات للتعامل مع العديد من الرسومات البيانية ذات الاتجاهين.

أمثلة عن الأوامر : Example Commands

```
.histogram y, frequency
```

يقوم هذا الأمر برسم مضلع تكراري للمتغير *y* موضحاً التكرارات على المحور العمودي.

```
.histogram y, start(0) width(10) norm fraction
```

يقوم برسم مضلع تكراري للمتغير *y* بعرض 10 وحدات للمضلع التكراري ويبدأ من نقطة 0، وإضافة منحنى التوزيع الطبيعي بناءً على المتوسط والانحراف المعياري للعينة، كما أن الرسم يعرض كسور البيانات على المحور العمودي.

```
.histogram y, by(x, total) percent
```

في شكل واحد يتم رسم عدة مضلعات تكرارية للمتغير *y* لكل قيمة من المتغير *x*، كما أن المضلع التكراري الأخير يعطي المجموع للعينة ككل ويعرض نسباً مئوية على المحور العمودي.

```
.kdensity x, generate(xpoints xdensity)
 width(20) biweight
```

يقوم هذا الأمر بإنشاء منحني الكثافة اللبي kernel density لتوزيع المتغير *x*، وإنشاء متغيرين جديدين في البيانات هما المتغير *xpoints* ويحتوي على

قيم المتغير x والتي تم تقدير كثافتها، والمتغير *xdensity* مع كثافة تم تقديرها ذاتياً، الخيار **width(20)** يحدد نصف عرض مركز الكثافة بالوحدات للمتغير x وإذا لم يتم تحديد الخيار **width( )** فإن الوضع الافتراضي هو اتباع صيغة بسيطة للمستوى الأمثل، أما الخيار **biweight** ففي مثالنا هذا سوف يحدد مركز كثافة ثنائي الوزن بدلاً من الوضع الافتراضي وهو استخدام نواة **Epanechnikov**.

```
.graph twoway scatter y x
```

يعرض رسم بياني للانحدار الخطي للمتغير y على المتغير x، الخيار **graph** هو جزء اختياري لجميع أوامر **twoway** فمثلاً يمكننا طباعة الأمر **twoway scatter** y x.

```
.graph twoway lfit y x || scatter y x
```

يقوم الأمر بإنشاء رسم بياني ثنائي لانحدار الخطي للمتغير y على المتغير x، حيث يظهر في الرسم خط الانحدار (خط الانحدار أو **lfit**) للمتغير y مقابل المتغير x كما يعرض الرسم أيضاً انتشار قيم المتغيرين x وy والذي يوضح أن درجة الثقة 95% في نطاق الانحدار الخطي.

```
.graph twoway scatter y x, xlabel(0(10)100)
 ylabel(-3(1)6,horizontal)
```

يقوم بإنشاء رسم بياني نقطي لانشاء قيم المتغير y مقابل المتغير x مع إعطاء المحور الأفقي x قيم 0، 10، ... 100 والمحور العمودي y قيم −3، −2، ... 6، مع كتابة الأسماء في وضع أفقي بدلاً من الوضع الافتراضي وهو الوضع العمودي.

```
.graph twoway scatter y x, mlabel(country)
```

يقوم بإنشاء رسم بياني نقطي لإنشاء قيم المتغير y مقابل المتغير x حيث يتم عرض البيانات على شكل نقاط، وكل نقطة لها اسم، وهو قيمة المتغير *country*.

```
.graph twoway scatter y x1, by(x2)
```

يقوم الأمر بإنشاء رسم بياني موحد لشكل انتشار المتغير y مقابل $x1$ ووضع مخطط منفصل لكل قيمة من قيم $x2$.

```
.graph twoway scatter y x1 [fweight =
 population], msymbol(oh)
```

يقوم الأمر بإنشاء رسم بياني نقطي لانتشار قيم المتغير *y* مقابل المتغير *x1* على شكل دوائر صغيرة (Oh) وتكون مواقع هذه الدوائر تميل للوزن التكراري للمتغير *population*

```
.graph twoway connect y time
```

يقوم هذا الأمر بإنشاء رسم بياني مبسّط للمتغير *y* مقابل المتغير *time* حيث يتم تمثيل البيانات بنقاط يصل بين كل نقطة وأخرى خط، ولعرض الخط الواصل بين هذه النقاط بدون إظهار النقاط قمنا باستخدام الخيار **line** بدلاً من الخيار **connect** كما يلي:

```
.graph twoway line y time
```

الطريقة الأخرى لإنشاء رسم بياني مبسط للفترة الزمنية، تتم باستخدام الأمر **tsline** والذي يعمل مع الأمر **tsset** وهذا الأخير يُستخدم لتحديد المتغير الذي يمثل الزمن (انظر الفصل 12).

```
.tsline y
.graph twoway line y1 y2 time
```

يقوم الأمر برسم بياني للفترة الزمنية (في هذا المثال الرسم البياني سوف يكون من النوع الخطّي) مع متغيرين هما *y1* و*y2* ولهما نفس المقياس، وسوف يتم تمثيلهما برسم بياني مع المتغير *time*.

```
.graph twoway line y1 time, yaxis(1) || line y2
 time, yaxis(2)
```

يقوم الأمر برسم بيانات المتغيرين *y1* و*y2* ولهما قياسات مختلفة، حيث يتم وضع المنحنيين على نفس الرسم البياني، ويكون المحور العمودي في جهة اليسار **yaxis(1)** يمثل قياس المتغير *y1*، بينما المحور العمودي في جهة اليمين **yaxis(2)** يمثل قياس المتغير *y2*.

```
.graph twoway contour temperature y x
```

هذا الأمر يقوم بإنشاء رسم بياني ملون يعرض متغير *temperature* على شكل شرائح مع المتغيرين *x* و*y* حيث إن هذين المتغيرين يشبهان إحداثيات

المواقع، وهناك مجموعة من الخيارات للتحكم في تفاصيل الشكل البياني، مثل عدد الألوان وطرق إضافة ألوان أخرى.

```
.graph matrix x1 x2 x3 x4 y
```

يقوم هذا الأمر بإنشاء مصفوفة من الرسومات البيانية في شكل واحد يعرض كل أشكال الانتشار المحتملة الممكنة لكل المتغيرات الموجودة.

```
.graph box y1 y2 y3
```

يقوم هذا الأمر بإنشاء رسم بياني على شكل صندوق لكل متغير من المتغيرات *y1, y2, y3*.

```
.graph box y, over(x) yline(23)
```

يقوم هذا الأمر بإنشاء رسم بياني على شكل صندوق للمتغير *y* مع كل قيمة من قيم المتغير *x* ورسم خط أفقي عندما تكون قيمة $y = 23$

```
.graph pie a b c
```

يقوم بإنشاء رسم بياني دائري مقسم إلى أجزاء، يشير كل جزء إلى القيمة المتعلقة بكل متغير من المتغيرات *a, b, c* ويجب أن تقاس هذه المتغيرات بنفس وحدة القياس.

```
.graph bar (sum) a b c
```

يعرض هذا الأمر مجموع المتغيرات *a, b, c* بأعمدة متراصة جنباً إلى جنب، وللحصول على المتوسط بدلاً من المجموع نقوم بطباعة الأمر **graph bar (mean) *a b c*** وهناك خيارات أكثر تتضمن الحصول على الوسيط ونسب مئوية وأعداد وإحصاءات أخرى لكل متغير (هذه الخيارات مثل **collapse**).

```
.graph bar (mean) a, over(x)
```

يقوم هذا الأمر بإنشاء رسم بياني يعرض متوسط المتغير *a* مع كل قيمة من قيم المتغير *x*.

```
.graph bar (asis) a b c, over(x) stack
```

يقوم هذا الأمر برسم أعمدة لكل قيم المتغيرات *a, b, c* حيث يتم وضع قيمة كل متغير فوق الأخرى في عمود واحد لكل قيمة من قيم المتغير *x*.

```
.graph dot (median) y, over(x)
```

يقوم هذا الأمر برسم شكل نقطي حيث إن كل نقطة تمثل علامة مقياس أفقي لقيمة الوسيط للمتغير *y* عند كل قيمة من قيم المتغير *x*، الأمر **graph dot** يدعم نفس الخيارات الإحصائية التي يدعمها الأمر **graph bar** والأمر **collapse**

```
.qnorm y
```

يقوم هذا الأمر بإنشاء رسم بياني يوضح التوزيع الطبيعي لقيم المتغير *y* المتعلقة مع فئات التوزيع الطبيعي.

```
.rchart x1 x2 x3 x4 x5, connect(1)
```

يقوم هذا الأمر بإنشاء رسم مراقبة الجودة R موضحاً مدى القيم الممثلة بالمتغيرات من المتغير *x1* إلى المتغير *x5*، وللحصول على قائمة كاملة بالخيارات التي يوفرها برنامج ستاتا عن هذا الأمر، قم بطباعة **help qc**، والأمر أعلاه يمكن تطبيقه من خلال استخدام قوائم ستاتا، وذلك باختيار Graphics > Quality control.

خيارات الرسم البياني مثل التحكم في العناوين والتوصيفات وعلامات الاختيار على المحاور في الرسم من الخيارات الأكثر شيوعاً في جميع أنواع الرسومات البيانية ببرنامج ستاتا، كما أن التركيب المنطقي لأوامر ستاتا الخاصة بالرسوم البيانية ثابت من نوع لآخر، وهذه العناصر المشتركة تعتبر ميزة رئيسة تسهّل عملية إنشاء أي رسم بياني.

المدرج التكراري : Histograms

يعرض المدرج التكراري توزيع مقاييس المتغيرات، ويتم إنشاء هذا المخطط بالأمر **histogram**، فمثلاً بالرجوع إلى بيانات مثالنا السابق عن 194 دولة في الفصل (2) والذي يحتوي على بيانات مؤشرات التطور البشري والتي تم جمعها بواسطة الأمم المتحدة.

```
.use C:\data\Nations2.dta, clear
.describe
```

```
Contains data from c:\data\Nations2.dta
  obs:           194                          UN Human Development Indicators
 vars:            13                          2 Jul 2012 06:11
 size:        12,804

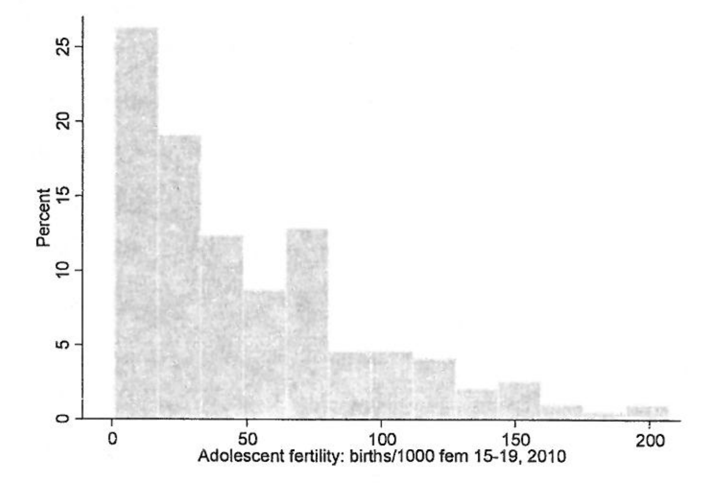
              storage  display    value
variable name   type   format     label      variable label

country         str21  %21s                  Country
region          byte   %8.0g      region     Region
gdp             float  %9.0g                 Gross domestic product per cap
                                               2005$, 2006/2009
school          float  %9.0g                 Mean years schooling (adults)
                                               2005/2010
adfert          float  %8.0g                 Adolescent fertility: births/1000
                                               fem 15-19, 2010
chldmort        float  %9.0g                 Prob dying before age 5/1000 live
                                               births 2005/2009
life            float  %9.0g                 Life expectancy at birth 2005/2010
pop             float  %9.0g                 Population 2005/2010
urban           float  %9.0g                 Percent population urban 2005/2010
femlab          float  %9.0g                 Female/male ratio in labor force
                                               2005/2009
literacy        float  %9.0g                 Adult literacy rate 2005/2009
co2             float  %9.0g                 Tons of CO2 emitted per cap
                                               2005/2006
gini            float  %9.0g                 Gini coef income inequality
                                               2005/2009

Sorted by:  region  country
```

الشكل (1.3) يعرض رسماً لمدرج تكراري بسيط للمتغير *adfert*، وهو يمثل معدل خصوبة المراهقين. وهذا الرسم البياني تم إنشاؤه بواسطة الأمر التالي:

```
.histogram adfert, percent
```



الشكل (1.3)

في قائمة ستاتا Edit > Preferences > Graph Preferences لـــدينا مجموعة من الخيارات لتصميمات معدة مسبقاً للألوان الافتراضية والظـــلال والرسومات البيانية، كما يمكن تحديد رسومات بيانية مخصصة لأغـــراض معينة، أغلب الأمثلة في هذا الكتاب سوف تستخدم لونين فقط مـــع هـــامش مظلل حول الرسم البياني. إن عملية اختيار اللونين الأبيض والأسود وألوان أخرى للرسومات البيانية يساعدنا في تحديد أفضل الأشكال لكل غرض مـــن الأغراض المطلوبة، والرسم البياني يتم إنشاؤه وحفظه تحت تصنيفات معينة، بحيث يمكن استعادته وإعادة حفظه تحت تصنيف آخر، كما سوف يتم الشرح لاحقاً.

خيارات الرسم يمكن استخدامها بعد إضافة فاصلة إلى أمر إنشاء الرسم البياني، فالرسم البياني بالشكل (1.3) يوضح خياراً واحدًا وهو النسبة المئوية **percent** (بدلاً من **density** وهو الخيار الافتراضي) حيث يتم عرض النسبة المئوية على المحور العمودي، وعند ظهور الرسم البياني على الشاشة توجد في نافذة الرسم مجموعة من قوائم الخيارات التي تتيح طباعة الرسم وحفظه ونسخه ولصقه في برامج أخرى مثل مايكروسوفت وورد.

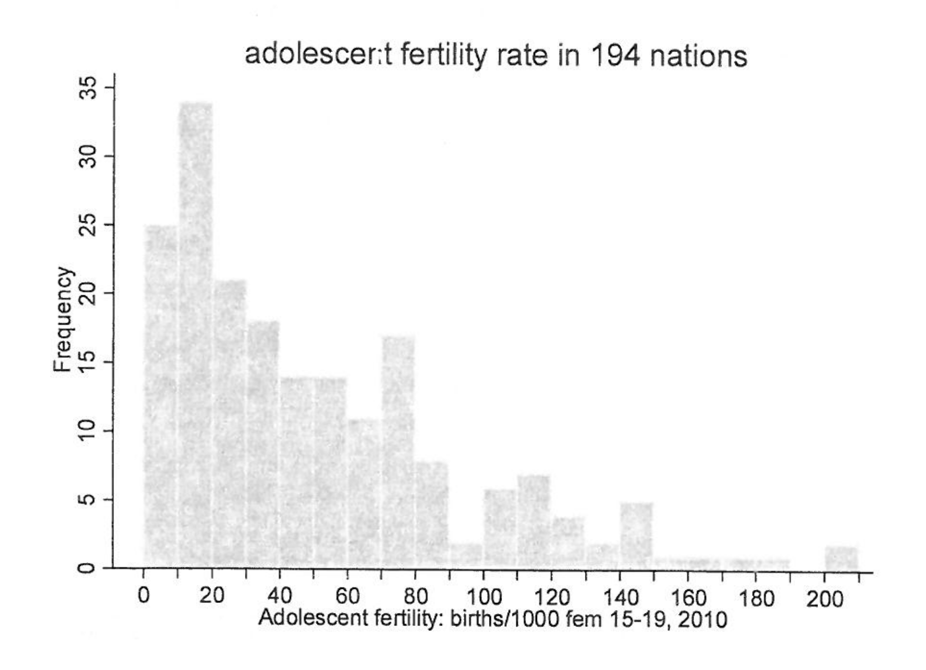
الشكل (1.3) يشير إلى التواء موجب للتوزيع مع منـــوال قريـــب مـــن الصفر، وحد أعلى حوالي 200، من الصعب إعطاء شرح أكثر دقة لهـــذا الرسم، لأن الأعمدة لم يتم رسمها لتعبر عن قيم المحور الأفقـــي x، الـــشكل (2.3) يعرض نفس الرسم البياني السابق بطريقة أخرى مع خيارات أخـــرى يمكن أن تجعل الرسم البياني أكثر وضوحاً:

frequency	التكرارات سوف تعرض على المحور العمودي y
start(0)	العمود الأول في المدرج التكراري سوف يبدأ من الصفر.
width(10)	عرض كل عمود في الرسم البياني سوف يكون 10 وحدات.
xlabel(0(20)200)	المحور الأفقي x سوف يبدأ من الصفر وحتى 200، بزيادة قـــدرها 20 بين كل قيمة وأخرى.
xtick(10(20)210)	سوف تكون هناك نقاط في المحور الأفقـــي x مـــن 10 إلـــى 210 بزيادة قدرها 20.

ylable(0(2)12,grid gmax)	المحور العمودي y سوف يبدأ من 0 إلى 35 بزيادة قدرها 5 وسوف يتم رسم خطوط شبكة أفقية تتضمن خطاً عند أعلى قيمة.
title("Adolescent fertility rate in 194 nations")	وضع عنوان للرسم البياني في أعلى الرسم.

الأمر أدناه والذي كُتِبَ في أربعة أسطر لتسهيل عملية قراءته، ولجعل هذه الأسطر تعمل في ملف تنفيذي do-file يمكننا إضافة ثلاث شرطات خلفية /// في نهاية كل سطر من الثلاثة أسطر الأولى، والتي تشير إلى أن الأمر مازال مستمراً في السطر التالي:

```
.histogram adfert, frequency start(0) width(10)
 xlabel(0(20)200) xtick(10(20)210)
 ylabel(0(5)35, grid gmax)
 title ("adolescent fertility rate in 194
 nations")
```

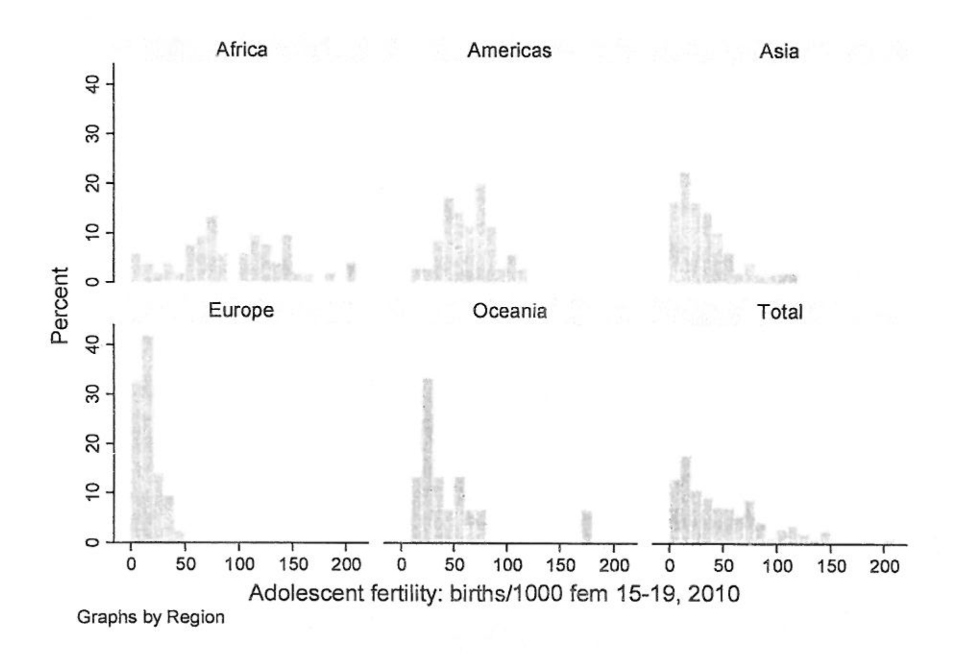


الشكل (2.3)

الشكل (2.3) يساعدنا في شرح التوزيع بشكل أكثر دقة، فمثلاً الآن يمكن أن نرى أن معدل خصوبة المراهقين في 34 دولة يتراوح بين 10 و20، وللحصول على قائمة كاملة بخيارات وتركيبة الأمر **histogram** قم بطباعة الأمر **help histogram**، وهناك أيضًا أمر آخر لرسم المدرج التكراري وهو الأمر **twoway histogram** حيث يسمح لنا باستخدام خيارت أكثر شيوعاً مع الأمر **twoway** سوف نناقشها لاحقاً، ويمكن أن تتعلم أكثر حول هذه الخيارات بطباعة الأمر **help twoway histogram**.

أحد خيارات الأمر **histogram** والتي يستخدمها مع أنواع أخرى من أوامر الرسم بياني هي قدرته على إنشاء عدة رسوم بيانية لكل قيمة لمتغير معين باستخدام الخيار **by(***varname***)**، الشكل (3.3) يعرض مدرجاً تكرارياً للمتغير *adfert* لكل إقليم من الأقاليم الخمسة التي تم دراستها مع رسم بياني سادس للمجموع **(total)** يعرض التوزيع لكل الأقاليم.

```
.histogram adfert, percent start(0) width(10)
 by(region, total)
```

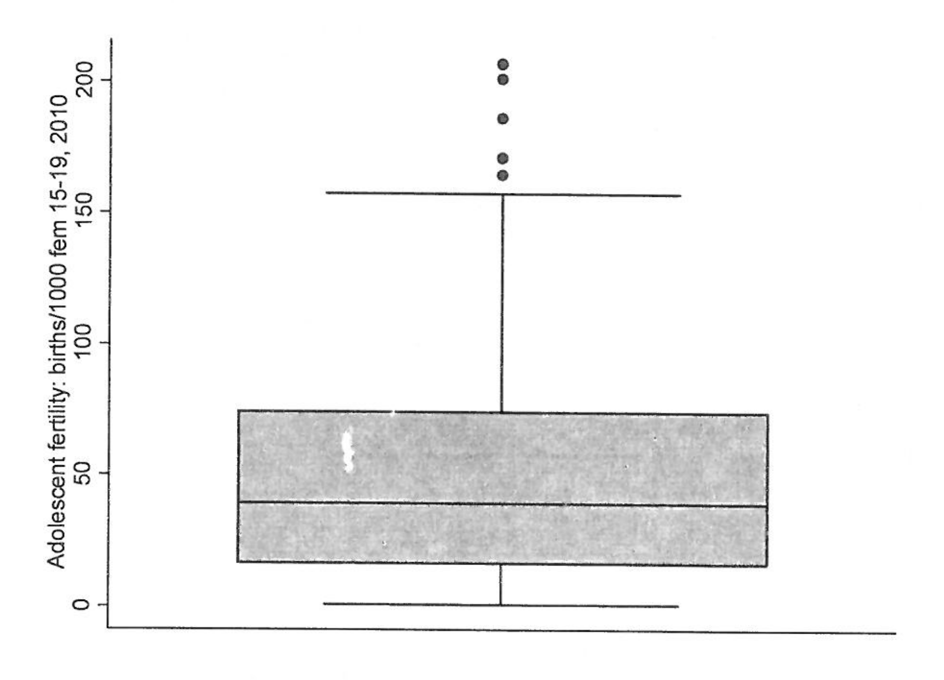


الشكل (3.3)

رسم الصندوق : Box Plots

رسم الصندوق يعطي نظرة سريعة عن مدى تركّز البيانات حول المركز وتشتتها وتناسقها وقيمها المتطرفة، فمثلاً الشكل (4.3) يعرض رسم صندوق مبسط للمتغير *adfert* (معدل خصوبة المراهقين)، حيث يمكن الحصول على هذا الرسم بطباعة الأمر التالي:

```
.graph box adfer
```



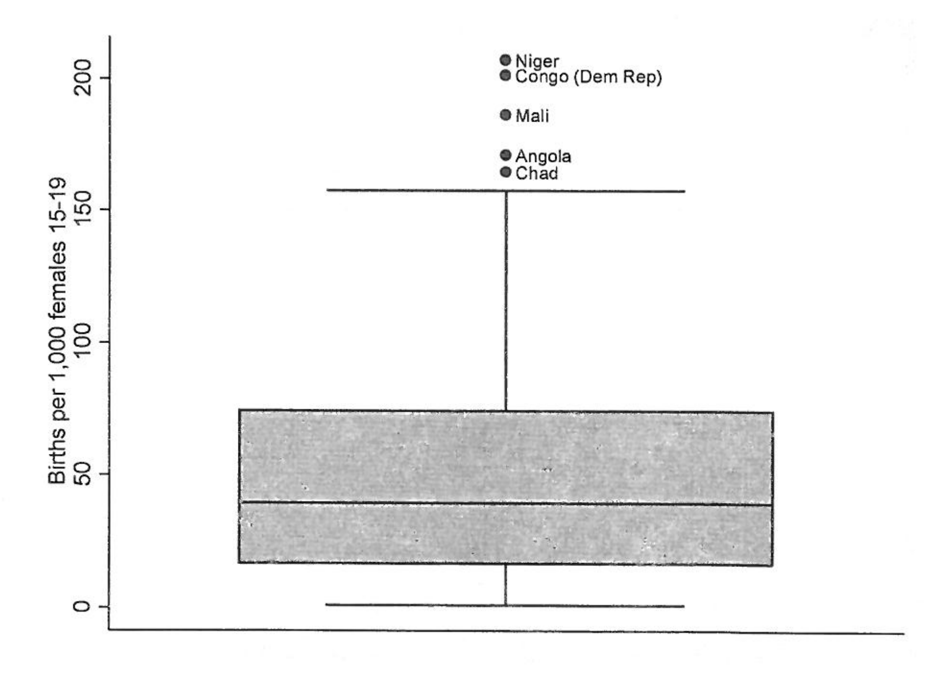
الشكل (4.3)

الشكل (4.3) يؤكد الالتواء الإيجابي لهذا التوزيع، كما يعرض أعلى خمس قيم متطرفة، والصندوق في الرسم يمتد من حوالي الربع الأول إلى الربع الثالث، وهذه المسافة تسمى المدى الربيعي interquartile range (IQR) ويتضمن المدى 50% تقريباً من البيانات، (رسم الصندوق في برنامج ستاتا يحدد الربيعات بنفس الطريقة في الأمر **summarize** والأمر **detail**) القيم المتطرفة تُعرّف بأنها قيم المشاهدات الأكبر من المدى الربيعي (5.1) وهي

تقع في النطاق الخارج عن المنطقة ما بين الربع الأول والربع الثالث، وقد تم توضيح القيم المتطرفة في الرسم باستخدام نقاط في أعلى الشكل البياني.

الشكل (5.3) يحدد القيم المتطرفة بالمتغير *adfert* ويصف نقاطها في الرسم مع قيم المتغير *country* (أسماء الدول)، كما يحدد عنواناً للمحور العمودي *y*، أما الخيار **marker** يمكنه أن يتحكم في الرموز والخصائص الأخرى التي تشير إلى القيم المتطرفة، فالخيار **marker(1)** في هذا المثال يشير إلى أول اسم في المتغير *y*، وهناك متغير واحد فقط باسم *y* ولكن في بعض الحالات يمكننا أن نجد اثنين أو أكثر، ونحتاج إلى وضع علامات لقيمها المتطرفة بشكل يوضح كل متغير على حدة.

```
.graph box adfert, marker(1, mlabel(country))
 ytitle("Births per 1,000 females 15-19")
```

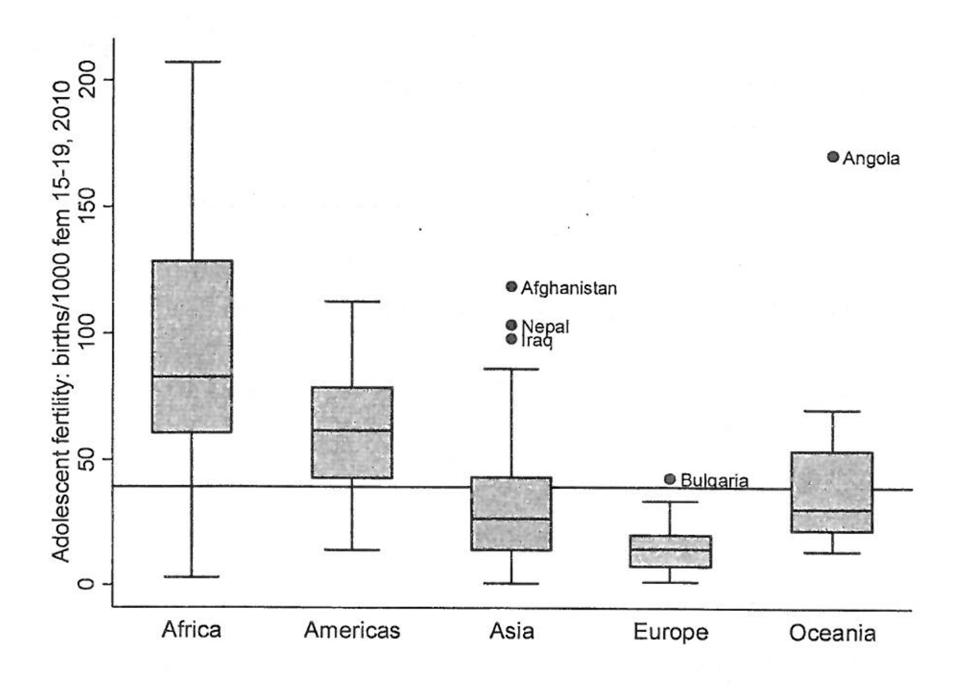


الشكل (5.3)

أحد أهم تطبيقات رسم الصندوق والأكثر شيوعاً هو مقارنة توزيع متغير ما مع فئات متغير ثان، الشكل (6.3) يقارن بين توزيع المتغير *adfert*

لمجموعة من الأقاليم *region*. وبصفة عامة، فإن الوسيط تمت الإشارة إليـــه بخط أفقي بواسطة الخيار **yline(39.3)**

```
.graph box adfert, marker(1, mlabel(country))
 yline(39.3) over(region)
```

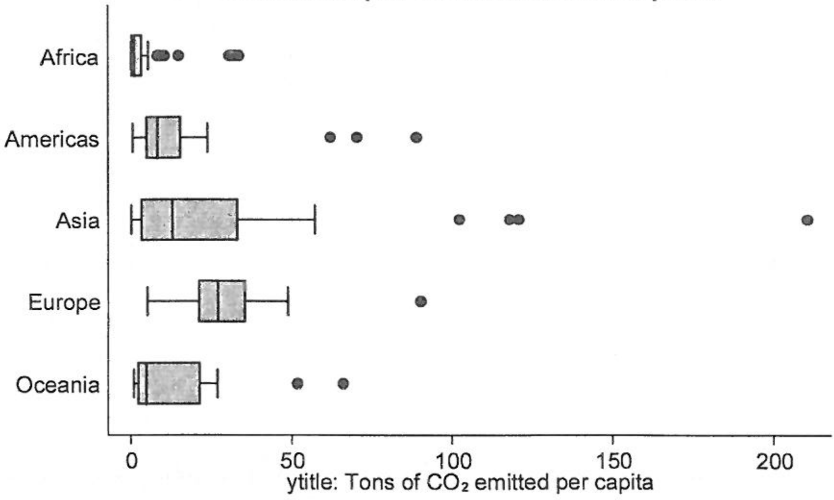


الشكل (6.3)

رسوم الصندوق يمكن أن يكون اتجاهها أفقياً بدلاً من الاتجاه العمودي، وذلك باستخدام الأمر **graph hbox**. الشكل (7.3) يستخدم معدل انبعاث ثاني أكسيد الكربون لكل فرد (*co2*) وهو متغير آخر موجود في ملـــف البيانــــات *Nations2.dta*، هذا المثال يعــرض مجمـــوعة مـــن خيـــارات العنـــاوين والتوصيفات التي يمكن تطبيقهـــا على أي نوع من الرســـومات البيانيــــة، فالخيار **note( )** والخيار **caption( )** تضع النص تحت الرسم البيـــاني، فـــي الشكل أدناه هناك عبارة "Statistics with Stata" تظهر بخط أســـود عـــريض وعبارة "Example of horizontal box plots" بخط مائـــل، وعنـــوان المحـــور العمودي، في الخيار **ytitle** (عنوان المحور العمودي *y* يظهر فيه كل رســـم

صندوق أفقي يشير إلى محور أفقي)، أما نسبة CO_2 فسوف يُشار إليها بطريقة مناسبة في الرسم، وتنسيق الخطوط في الرسم البياني يمكن التحكم به داخل الرسم نفسه باستخدام خاصية علامات ستاتا ولغة التحكم Stata markup and control language (SMCL)، لمزيد من الأمثلة عن ذلك قم بطباعة الأمر **help graph text**

```
.graph hbox co2, over(region)
 note("note: {bf:Statistics with Stata},
 version 12")
 caption("c aption: United Nations Human
 Development Report
 2011")
 title("title: {it:Example of horizont al box
 plots}")
 ytitle("ytitle: Tons of CO{subscript:2}
 emitted per capita")
```



الشكل (7.3)

القيم المتطرفة الفردية لم يتم توصيفها في الشكل (7.3) لأنه من الصعب قراءة الوصف في التنسيق الأفقي. أما القيم المتطرفة الثلاث في أمريكا Americas هي الولايات المتحدة وكندا وجزيرة الترينيداد وتوباجو (وهي جزيرة صغيرة في شمال أمريكا الشمالية تشتهربصناعة النفط والغار في الكاريبي)، ونرى أن أعلى معدل لإنبعاث ثاني أكسيد الكربون للفرد بدول القارة الأسترالية في أستراليا Oceania، وهناك أربع دول مصدرة للنفط فيها أعلى قيم متطرفة في آسيا Asia؛ وبإلقاء نظرة أكثر تفحصاً للقيم المتطرفة، فإن رسم الصناديق يوضح أننا أحياناً نجد أن هناك بعض المشاهدات المثيرة في البيانات نفسها وليس فقط في الحسابات الإحصائية.

هناك عدد كبير من خيارات التحكم في المظهر والظلال وتفاصيل الصناديق في الرسم البياني، وللحصول على قائمة بهذه الخيارات قم بطباعة الأمر **help graph box**، وصف المحاور والنقاط والعناوين يمكن تعديلها بخيارات **by(***varname***)** أو **by(***varname*, **total)** وهذه الخيارات تعمل بنفس الطريقة مع أوامر رسم بياني أخرى ببرنامج ستاتا، فمثلاً **by(***region***)** سوف يقوم بإنشاء رسم صندوق في خمس نوافذ صغيرة بدلاً من خمسة صناديق في رسم بياني **over(***region***)** واحد كما تم القيام به في الشكل (6.3) والشكل (7.3).

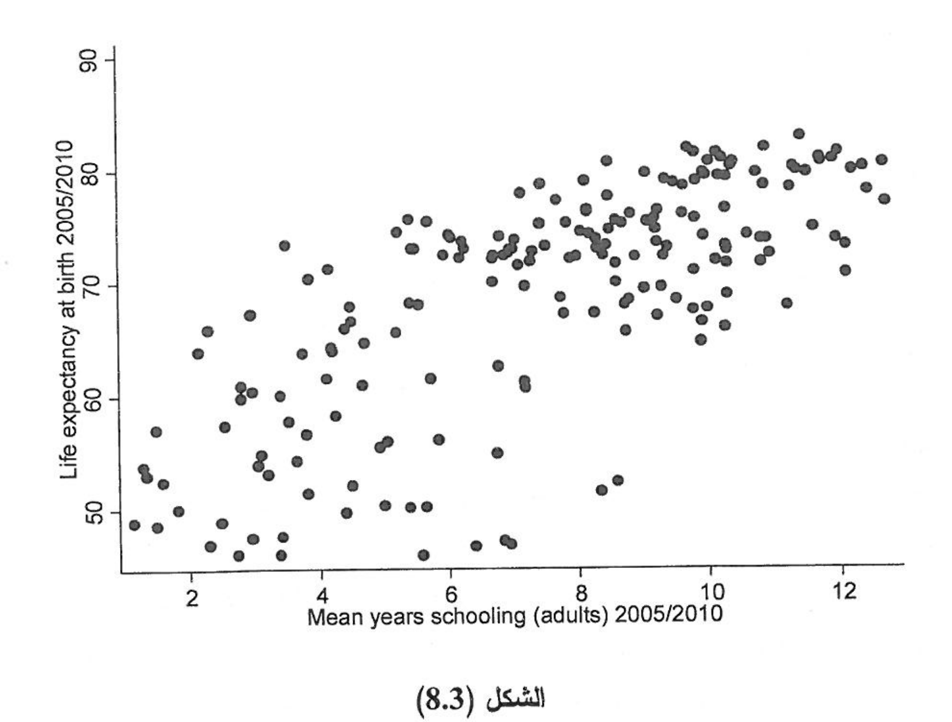
شكل الانتشار وتركيباته : Scatterplots and Overlays

شكل الانتشار هو جزء من عدد كبير من الرسومات البيانية التي تُسمى **twoway**، والأمر الأساسي لإنشاء شكل الانتشار هو:

```
.graph twoway scatter y x
```

حيث إن المتغير *y* يمثل المحور العمودي، والمتغير *x* يمثل المحور الأفقي (الجزء الرئيس للأمر **graph twoway** اختياري ولكن تم كتابته هنا للتأكيد على أهميته التي سوف تتضح لاحقاً) فمثلاً باستخدام بيانات الملف *Nations2.dta* يمكننا إنشاء رسم بياني للمتغير *life* (متوسط العمر المتوقع) مع المتغير *school* (متوسط سنوات الدراسة) وصورة هذا الرسم البياني تم عرضها في الشكل (8.3) حيث إن كل نقطة في الشكل تمثل دولة واحدة.

```
.graph twoway scatter life school
```



الشكل (8.3)

كما رأينا سابقاً في المضلع التكراري، فإنه يمكننا استخدام العديد من الخيارات مثل **xlabel, xtitle, ylabel** للتحكم في توصيف المحور الأفقي والمحور العمودي والعناوين.. الخ، وكما رأينا في رسم الصندوق، فإن شكل الانتشار يسمح لنا بالتحكم في شكل الرسم ولونه وحجمه وتوصيفه وسمات نقاطه، الشكل (8.3) يستخدم النقاط الافتراضية في الرسم وهي عبارة عن دوائر غامقة، وهناك بعض التأثيرات التي يمكننا تضمينها في الرسم باستخدام الخيار **msymbol(O)**، كما يمكننا استخدام خيارات أخرى مثل **msymbol(Th)** (مثلثات مجوفة كبيرة)، **msymbol(d)** (علامات ماسيّة صغيرة)، **msymbol(+)** (علامات زائد)، **msymbol(i)** (رموز مخفية، وهي طريقة عملية لبعض الأغراض الخاصة)، وللحصول على قائمة كاملة بالعلامات والخيارات قم بطباعة الأمر **help scatter**.

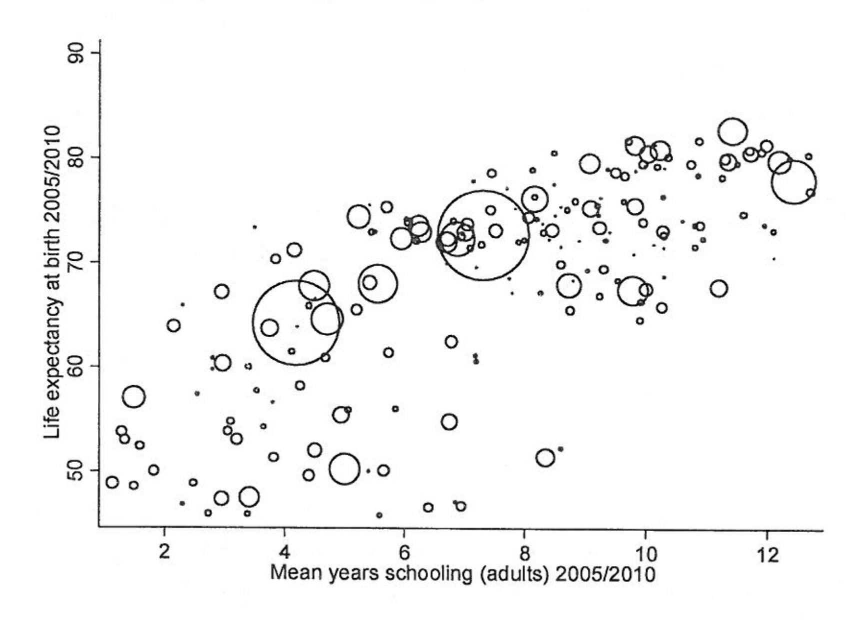
الخيار **mcolor** يتحكم في لون النقاط في الرسم البياني، فمثلاً الأمر:

```
.graph twoway scatter waste metro, msymbol(S)
 mcolor(purple)
```

سوف يقوم بإنشاء شكل الانتشار، بحيث إن كل نقطة بالرسم عبارة عن مربع أرجواني كبير، وللحصول على قائمة كاملة بالألوان المتـوافرة، قـم بطباعة الأمر **help colorstyle** وهذه الألوان قابلة للتطبيـق علـى الأعمـدة والخطوط والنصوص والعناصر الأخرى في أي رسم بياني ببرنامج ستاتا.

أحد الخيارات المثيرة التي يمكن القيام بها مع شكل الانتشار هو جعـل حجم الرمز متناسباً مع متغير ثالث، حيث يتم إعطاء نقاط البيانـات شـكلاً مختلفاً. فعلى سبيل المثال، قد نقوم بإنشاء شكل انتشار للمتغير *life* والمتغير *school* ولكن حجم النقاط بالرسم البياني سوف يعكس عدد السكان فـي كـل دولة (*pop*)، وهذا ما تـم عملـه فـي الـشكل (9.3) باسـتخدام الخيـار [**fweight=***varname*] او باستخدام خاصية الوزن التكراري، كمـا أن خيـار الدوائر المجوفة **msymbol(Oh)** يعتبر شكلاً مناسباً أكثر.

```
.graph twoway scatter life school
 [fweight=pop], msymbol(Oh)
```



الشكل (9.3)

الأوزان التكرارية مفيدة مع بعض أنواع الرسم البياني الأخـرى. وزن بعض العناصر يمكن أن يكون موضوعاً غامضاً ومعقدًا، لأن الأوزان يمكن تمثيلها بأنواع مختلفة، ولها معاني مختلفة حسب السياق، وللحصول على معلومات عامة عن الأوزان يمكنك طباعة الأمر **help weight**.

الخاصية الرئيسة لمجموعة أوامر **graph twoway** هي قدرتها على وضع شكلين أو أكثر معاً لإنشاء أشكال أكثر تعقيداً، فمثلاً لإنشاء شكل انتشار للمتغير *life* مع المتغير *school* باستخدام دوائر مجوفة كعلامات بالرسم، يمكن ذلك بطباعة الأمر

```
.graph twoway scatter life school, msymbol(Oh)
```

خطوط الانحدار البسيطة (**lfit**) هي نوع مختلف من رسومات **twoway**، ولمشاهدة خط انحدار المتغير *life* على المتغير *school* مع تنسيق خط الانحدار ليكون الخط عريضاً بدرجة متوسطة نقوم بطباعة الأمر.

```
.lfit life school, lwidth(medthick)
```

ولكن عادة نريد أن نرى شكل الانتشار، وخط الانحدار معاً، يمكن القيام بذلك عن طريق وضع الرسم البياني **lfit** الخاص بخط الانحدار فوق شكل الانتشار **scatter** باستخدام أمر واحد مع || ("أنابيب") تشير إلى أنه تم وضع الشكلين معاً، الأمر أدناه تم كتابته في سطرين ولكن يُفترض أن تتم طباعته في سطر واحد.

```
.graph twoway scatter life school, msymbol(Oh)
 || lfit life school, lwidth(medthick)
```

أخيراً، إذا كان لدينا خيارات مؤكدة يمكن تطبيقها على الرسم البياني ككل، فإن هذه الخيارات يمكن وضعها في نهاية الأمر بعد ||، الشكل (10.3) يوضح ذلك، فالخيارت العادة لا تتضمن فقط **ylabel, xlabel, xtick** ولكن تحدد أيضاً بعض التفاصيل حول مربع شرح الرسم البياني **legend**.

```
.graph twoway scatter life school, msymbol(Oh)
 || lfit life school, lwidth(medthick)
 || , ylabel(45(5)85) xlabel(2(2)12)
 xtick(1(2)13)
 legend(col(1) ring(0) position(11))
```



الشكل (10.3)

خيار **legend** في الشكل (10.3) يحدد ثلاثة عناصر هي:

col(1) مربع شرح الرسم يُفترض أن يكون في عمود واحد، وبالتالي سوف يظهر في صفين اثنين.

ring(0) مربع شرح الرسم يجب أن يقع داخل منطقة الرسم بدلاً من خارجها، فعندما يقع مربع شرح الرسم في خارج الرسم يؤدي إلى ازدحام لعناصر البيانات في مساحة صغيرة.

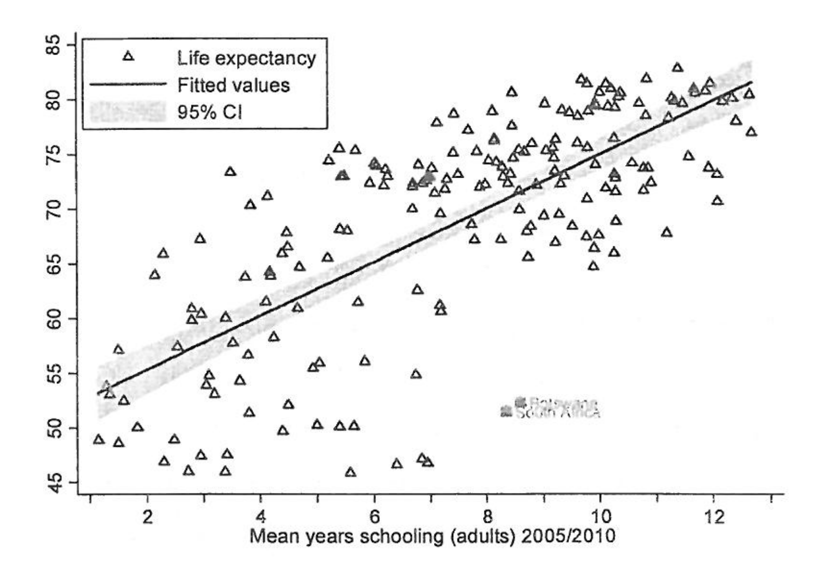
position(11) وهذا الأمر يضع مربع شرح الرسم في موقع الساعة 11 وهذا ما حدث في الرسم أعلاه حيث كان هذا الموضع خال من أي بيانات.

عملية وضع مربع شرح الرسم داخل الرسم نفسه في جهة خالية من أي بيانات، يُعتبر أفضل طريقة إذا كنا نستطيع التحكم في ذلك، ويمكنك اختيار ذلك من خلال وضع مربع شرح الرسم حسب الاختيار الافتراضي لبرنامج ستاتا أو تغيير موقعه إلى مكان آخر حسب ما تراه مناسباً، ويمكنك الاطلاع على خيارات مربع شرح الرسم **legend_options** بطباعة الأمر **help twoway**

options وذلك للتعرّف على الخيارات الممكنة للتحكم في موقع مربع شــرح الرسم ومحتوياته وكيفية ظهوره.

الشكل (11.3) يحاول التعامل مع هذه الأفكار بوضع ثلاثة أشكال مــع بعضها. أحد هذه الأشكال الثلاثة من نوع **twoway**. الخيار **lfitci** يعني رسم الانحدار الخطي مع فترة ثقة، وللقيام بذلك علينا أولاً تحديد **lfitci** ثــم يــتم وضع شكلي الانتشار فوق خط الانحداروبهذا سوف نرى نقاطاً فوق فتــرات الثقة الرمادية، إذا قمنا بتحديد **lfitci** في آخر سطر الأمر، فإن فترات الثقــة سوف يتم رسمها فوق بعض النقاط في الرسم البياني النقطي.

```
.graph twoway lfitci life school,
 lwidth(medthick)
 || scatter life school, msymbol(Th)
 || scatter life school if school>8 & life<55,
 msymbol(S) mlabel(country)
 ||, ylabel(45(5)85) xlabel(2(2)12)
 xtick(1(2)13)
 legend(col(1) ring(0) position(11) label(3
 "Life expectancy")
 order(3 2 1))
```



الشكل (11.3)

الخيار الافتراضي lfitci: يعرض فترة الثقة للمتوسط الحسابي الشرطي *conditional mean* للمتغير *y* بدلاً من عرضه للقيم الفردية المتوقعة، وسوف يشير برنامج ستاتا إلى الخطأ المعياري للمتوسط الشرطي بعبارة "الانحراف المعياري للتنبؤ" أو **stdp**. إذن فالخيار الافتراضي في الشكل (11.3) يعادل طباعة الأمر.

```
.graph twoway lfitci life school, stdp
```

الأخطاء المعيارية للقيم الفردية المتوقعة يمكن وصفها بأنها "الانحراف المعياري للتنبؤ" أو **stdf**، ولمشاهدة عرض أكبر لفترات الثقة للتوقعات الفردية نقوم طباعة الأمر:

```
.graph twoway lfitci life school, stdf
```

الشكلان الآخران في الشكل (11.3) كلاهما رسومات بيانية لشكل الانتشار توضح كيف يتم وصف (أو إنشاء رسم بياني مع رموز مختلفة) مشاهدات معينة، وقد تم تحديد قيمتين متطرفتين عن طريق إنشاء شكل انتشار عادي مع كل البيانات، واختيار مثلثات مجوفة كعلامات لهذه النقاط.

```
|| scatter life school, msymbol(Th)
```

ثم نقوم بوضع الشكل مع شكل الانتشار الثاني (وهو الرسم البياني الثالث في الصورة أعلاه) مع استخدام المحدد **if** لتحديد الدول التي فيها متوسط التعليم أكبر من 8 سنوات، ومتوسط العمر المتوقع أكبر من 55 سنة، هناك دولتان فقط في أسفل اليمين تقابل هذه المعايير، وتم الإشارة إليهما بمربعات غامقة وتم توصيفهما بأسماء الدول وهما بوتسوانا Botswana وجنوب أفريقيا South Africa فهاتان الدولتان اللتان يتمتعان بمستوى تعليم عالي ولكن مع متوسط عمر متوقع منخفض الأمر الذي جعلهما يبرزان عن بقية الدول الأخرى.

```
|| scatter life school if school>8 &life<55,
msymbol(S) mlabel(contry)
```

الخيارات العامة في الشكل (11.3) تحدد توصيفات المحور الأفقي *x* وعلامات الرسم، كما تتحكم في مربع شرح الرسم، ومرة أخرى نعطي

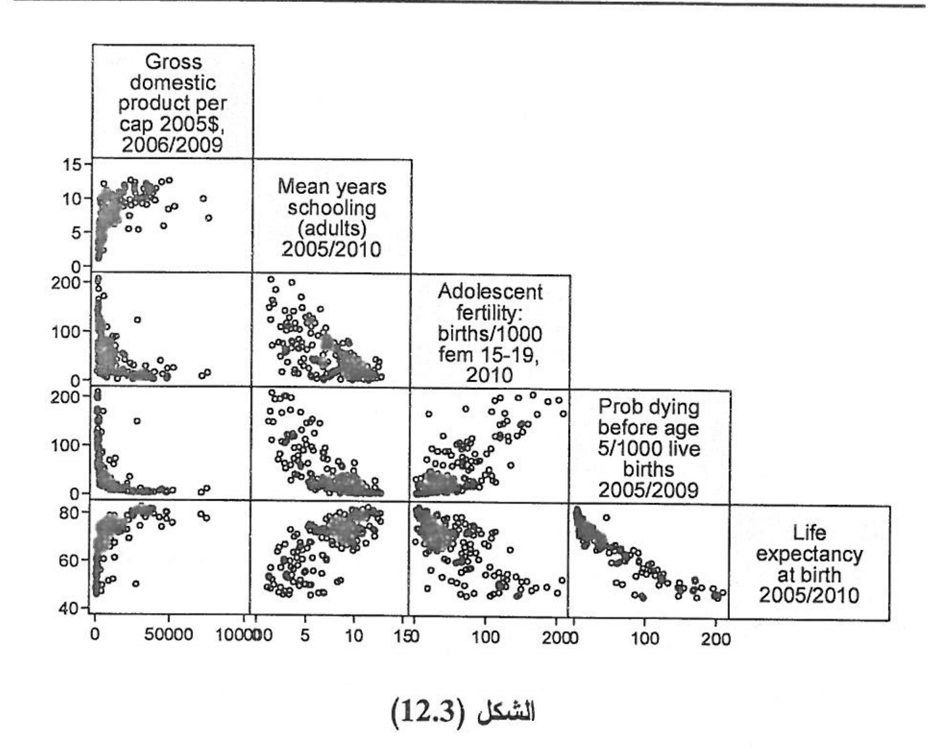
لمربع شرح الرسم عمودًا واحدًا، ونضعه في موقع الساعة 11 داخل الرسم نفسه، وتوصيف المتغير الثالث *y* في مربع شرح الرسم هو "متوسط العمر المتوقع" بدلاً من التوصيف الأطول والذي يتم استخدامه كخيار افتراضي.

```
legend(col(1) ring(0) position(11) label(3
"Life expectance")
```

الخيار legend : ينتهي بالخيار الفرعي **order(3 2 1)** الذي يحدد بأن مربع شرح الرسم يجب أن يتم ترتيبه 3–2–1، وهذا لايبدو بسيطاً كما يظهر لأن طريقة ستاتا في فهم الأشكال البيانية الثلاثة المتراكبة في الشكل (11.3) تتضمن في الواقع أربعة متغيرات يمكن وضعها في مربع شرح الرسم، وبإعطاء عدد متسلسل للمتغيرات في الأمر **graph twoway** هذا يعتبر (1) 95% في فترة الثقة، (2) القيم المناسبة أو توقعات الانحدار الخطي، (3) معدل العمر المتوقع في كل الدول للرسم البياني الأول وهو شكل الانتشار، (4) معدل العمر المتوقع لكل من بوتسوانا Botswana وجنوب أفريقيا South Africa في شكل الانتشار الثاني، بواسطة تحديد الخيار **order(3 2 1)** فإننا نحدد بأن يكون هناك مربع لشرح الرسم البياني في الشكل (11.3) وهذا المربع يقوم بترتيب (3) أولاً و(3) ثانياً و(1) أخيراً، تاركاً الشكل الرابع في الخارج حيث لم تتم الإشارة إليه في الخيار الفرعي **order()**، لذا فإن الخيار الفرعي **order()** لايتحكم فقط في ترتيب ظهور المتغيرات في مربع شرح الرسم ولكن يتحكم أيضاً في ظهورها.

مصفوفات شكل الانتشار في الرسم البياني لشكل الانتشار ليست من نوع **twoway** ولا يمكن تركيبها على رسومات أخرى، ولكنها تحتوي على رسومات بيانية متعددة يمكن أن تأخذ نفس الرموز النقطية في الرسم، الشكل (12.3) يعرض مصفوفة شكل الانتشار لرسومات بيانية تمثل شكل الانتشار لخمسة متغيرات في ملف البيانات *Nations2.dta*

```
.graph matrix gdp school adfert chldmort life,
 half msymbol(oh)
```



الشكل (12.3)

مصفوفة الانتشار في الرسم البياني لشكل الانتشار تمثل صورة مرئية لمصفوفة الارتباط والتي تعتبر مفيدة في حالة تحليل المتغيرات المتعددة، حيث إنها تعرض ملخصاً للعلاقة بين عدد المتغيرات المتشابهة، مما يسمح بملاحظة وجود أي علاقات غير خطية أو قيم متطرفة أو الميول العنقودي الذي قد يؤثر على النماذج الإحصائية. العلاقات غير الخطية تتضمن متغير الناتج المحلي الإجمالي للفرد *gdp* والذي يظهر بوضوح في الشكل (12.3) معطياً تحذيرات حول هذا الميول حيث لا يمكننا رؤية هذا الميول في مصفوفة الارتباط وحدها.

الخيار **half** يحدد بأن الشكل (12.3) يجب أن يحتوي على الجزء الثلاثي السفلي للمصفوفة فقط، أما الجزء الثلاثي العلوي فسوف يتم إهماله لتماثله التام مع الجزء السفلي، الخيار **msymbol(o)** يُسمى الدوائر الصغيرة، ويُستخدم لوضع دوائر صغيرة كعلامات في شكل الانتشار، ويجب ملاحظة أن التحكم في المحاور في هذا النوع من الأشكال، يعتبر أكثر تعقيداً بسبب

وجود العديد من المتغيرات والمحاور. ولمزيد من التفاصيل عن التحكم في المحاور قم بطباعة الأمر **help graph matrix**.

عند احتواء المتغيرات على متغير مستقل واحد أو متغير مؤثرواحد من ناحية، وعدة متغيرات تابعة أو متغيرات متأثرة من ناحية أخرى، فإن هذا يساعد في عمل قائمة للمتغيرات التابعة التي ظهرت عند استخدام الأمر **graph matrix**، وهذا يؤدي إلى الحصول على رسم بياني متناسق للمتغيرات التابعة والمتغيرات المستقلة في الصف السفلي للمصفوفة، فالمتغير الأخير في الصف السفلي بالشكل (12.3) وهو متوسط العمر المتوقع *life* يمكن تحليله كمتغير تابع في الفصلين (7 و8).

الرسومات البيانية الخطيّة والخطية المتصلة :

Line Plots and Connected-Line Plots

الرسومات البيانية الخطية المتصلة **(graph twoway connect)** هي عبارة عن رسومات بيانية لشكل الانتشار تم وصل نقاطها بخطوط، أما الرسومات البيانية الخطية **(graph twoway line)** فتقوم بعرض خطوط بدون علامات مثلما رأينا في أشكال الانتشار. ولكن كلا النوعين ينتميان لنوع **graph twoway** والتي يمكن تركيبهما ودمجهما في رسم بياني واحد، خيارات أشكال الانتشار التي تتحكم في توصيف المحاور وعلاماتها تقوم بنفس الوظيفة في الرسومات الخطية، والرسومات الخطية المتصلة، وهناك خيارات أكثر للتحكم في عرض خط المحور ونمطه ولونه وعدة خصائص أخرى.

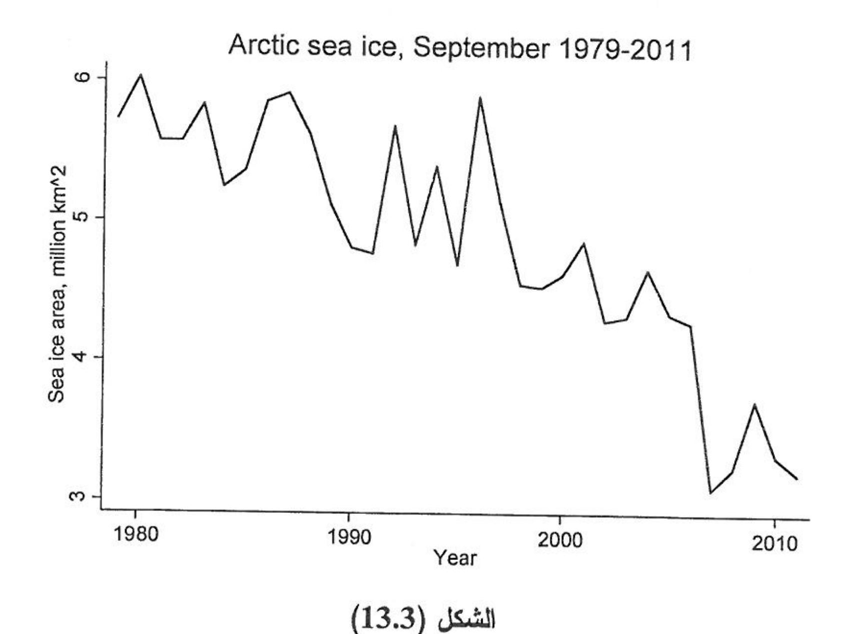
الرسومات البيانية الخطية والخطية المتصلة، لها استخدامات أكثر من أشكال الانتشار، فمثلاً يمكنها تمثيل التغيرات في متغير ما خلال فترة من الزمن. ولشرح هذه الرسومات البيانية، سوف نعود لاستخدام ملف البيانات *Arctic9.dta* والذي يحتوي على مشاهدات عن الجليد، ودرجات الحرارة في القطب الشمالي لفترة 33 سنة.

```
.use C:\data\arctic9.dta, clear
.describe
```

```
Contains data from c:\data\arctic9.dta
  obs:            33                          Arctic September mean sea ice 1979-2011
 vars:             8                          2 Jul 2012 06:11
 size:           891
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
year            int    %ty                    Year
month           byte   %8.0g                  Month
extent          float  %9.0g                  Sea ice extent, million km^2
area            float  %9.0g                  Sea ice area, million km^2
volume          float  %8.0g                  Sea ice volume, 1000 km^3
volumehi        float  %9.0g                  Volume + 1.35 (uncertainty)
volumelo        float  %9.0g                  Volume - 1.35 (uncertainty)
tempN           float  %9.0g                  Annual air temp anomaly 64N-90N C
-------------------------------------------------------------------------------
Sorted by:  year
```

الرسم البياني الخطي للمتغير *area* مع المتغير *year* يوضح الانخفاض في المنطقة الجليدية بالقطب الشمالي في شهر سبتمبر خلال الفترة، ولكن هناك انخفاضاً غير طبيعي في سنة 2007 على وجه الخصوص (انظر الشكل 13.3).

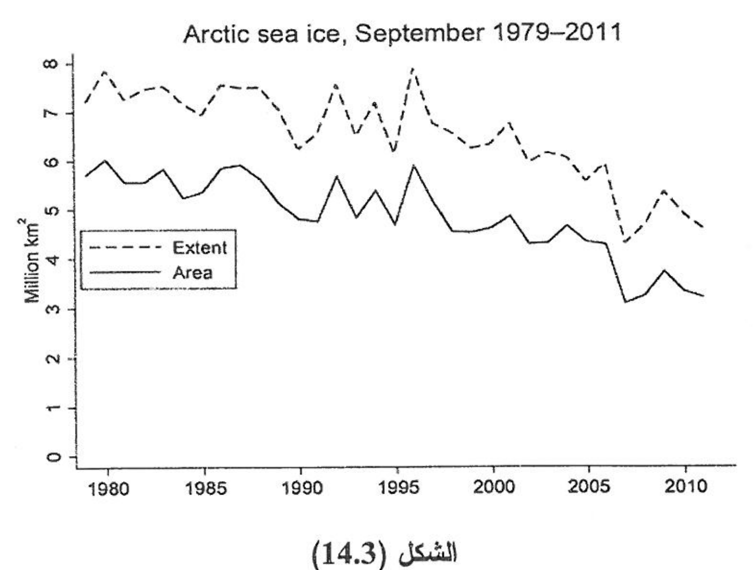
```
.graph twoway line area year,
title("Arctic sea ice, September 1979-2011")
```



الشكل (13.3)

الشكل (14.3) يوضح الفترة الزمنية بطريقة أكثر دقة مضيفاً خطاً ثانياً لامتداد الجليد بالبحر *extent* (المنطقة المغطاة بالجليد بنسبة 15% على الأقل)، كما تم توصيف المحور الأفقي *x* بالسنوات من 1980 إلى 2010 بفرق 5 سنوات بين كل سنة وأخرى **(xlable(1980(5)2010)** ولكن تم تقليص توصيف المحور الأفقي لأن متغير السنة واضح ولتوفير مساحة أكبر في الرسم، أما المحور العمودي *y* يبدأ من الصفر – وهو معدل محتمل وذو معنى لبعض المهتمين بالقطب الشمالي – إلى 8 بزيادة قدرها 1 مع خطوط شبكة في الرسم تتضمن أقل قيمة وهي 0 وأكبر قيمة وهي 8 وتم توصيف القيم بالأمر **ylabel(0(1)8, grid gmin gmax)**.

```
.graph twoway line area extent year,
xlabel(1980(5)2010)
xtitle(" ")
lwidth(medium medthick) lpattern(solid dash)
legend(row(2) ring(0) position(9)
label(1 "Area") label(2 "Extent") order(1 2))
ylabel(0(1)8, grid gmin gmax) ytitle("Million
km{superscript:2}")
title("Arctic sea ice, September
1979'=char(150) '2011")
```



الشكل (14.3)

المتغير الأول في الشكل (14.3) وهو *area* تم رسمه بخط ذي عرض متوسط **medium** ومتصل **solid**. المتغير الثاني وهو *extent* تم رسمه بخط أعرض قليلاً **medthick** ومتقطع **dash**، خيار **legend()** يحدد التوصيفات بمربع شرح الرسم ويضع المتغير *extent* أولاً بمربع شرح الرسم ويضع المربع بموقع أعلى في الرسم.

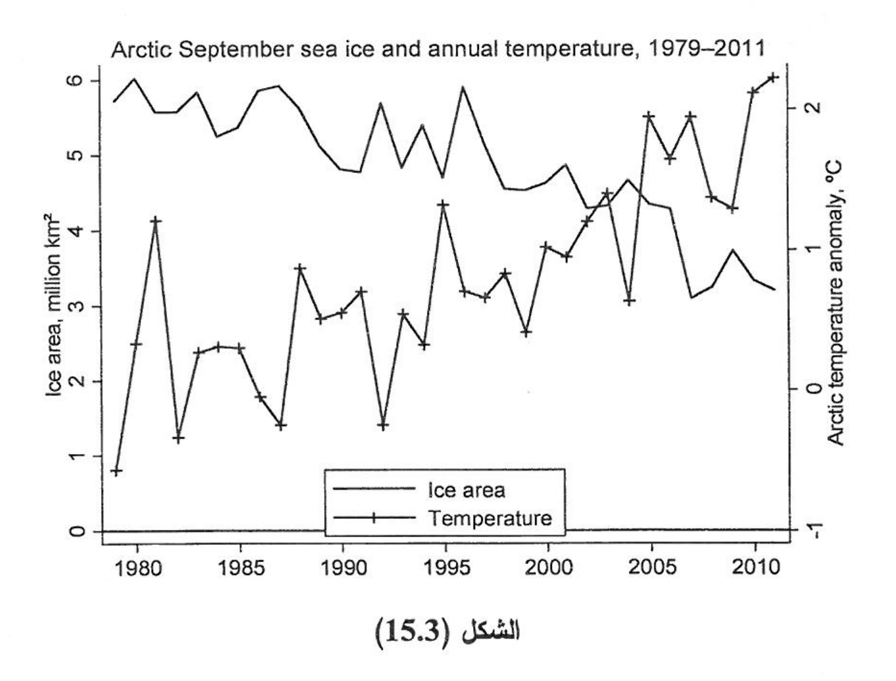
المنطقة الجليدية ومدى هذه المنطقة تم قياسه بمليون متر مربع، والخيار **ytitle("Million km{superscript:2}")** يعرض عنوان المحور العمودي "Million km^2"، وللحصول على تفاصيل أكثر حول التحكم في النصوص بالرسم البياني، قم بطباعة الأمر **help graph text**، كما أن الشكل (7.3) يعرض أمثلة أخرى عن ذلك.

طريقة أخرى لتغيير النصوص بالرسم البياني في الشكل (14.3) وهي مناسبة أكثر: السنوات 1979-2011 في العنوان يمكن فصلها بعلامة (–) بدلاً من علامة (-)، فالعلامة الأولى هي خاصية افتراضية لا تظهر في لوحة المفاتيح، ولكن يمكن تحديدها باستخدام رمز ASCII 150، خيار **title** في الشكل (14.3) تم إدراجه بواسطة ASCII والرمز 150 بين سنة 1979 وسنة 2011 بواسطة كتابة **1979`=char(150)'2011**، لاحظ أنه تم استخدام علامات التنصيص في يمين السنة الأولى ويسار السنة الثانية.

الشكل (15.3) يستخدم علامة (–) مع رمزين من رموز ASCII في الجزء الخاص بالوقت في الرسم البياني، حيث إن المتغيرين اللذين بالرسم البياني تم قياسهما بمقاييس مختلفة، ولكن تم وضعهما في رسم بياني واحد. فعلى سبيل المثال، قمنا بوضع منحنى لمتغير المنطقة *area* مع متغير السنة *year* والمنحنى عبارة عن خط متصل (الخيار **connect** يجمع بين خاصيتين هما **scatter** و**line**) لدرجة حرارة القطب الشمالي مع السنة *year*، كما أن الرسم البياني يوضح أن الجليد في القطب الشمالي يتناقص بسبب زيادة درجة حرارة الهواء في القطب الشمالي. المتغيران يؤثر كلاهما على الآخر، ويعكسان تغيرات أكبر ظهرت خارج منطقة القطب الشمالي.

```
.graph twoway line area year, ylabel(0(1)6)
yline(0)
ytitle("Ice area, million km`=char(178)'")
```

```
|| connect tempN year, yaxis(2) ylabel(-1(1)2,
axis(2)) msymbol(+)
ytitle("Arctic temperature anomaly,
`=char(186)'C", axis(2))
|| , xlabel(1980(5)2010) xtitle("")
legend(row(2) ring(0) position(6)
label(1 "Ice area") label(2 "Temperature")
order(1 2))
title("Arctic September sea ice and annual
temperature,
1979`=char(150)'2011", size(medium))
```



الشكل (15.3)

في الشكل (15.3) درجة الحرارة السنوية في القطب الشمالي تم تمثيلها على المحور العمودي الثاني، والذي يظهر في يمين الرسم البياني، وقيم درجات الحرارة تم تمثيلها بعلامات زائد ((msymbol(+)) أما مربع شرح الرسم تم وضعه في داخل الرسم البياني، وبالتحديد في موقع الساعة السادسة، موضحاً المتغيرين في الرسم، وبدلاً من استخدام {superscript:2} لكتابة km^2 في المحور العمودي بالجهة اليسرى (كما تم سابقاً في الشكل

14.3). أما الشكل (15.3) استخدم رموزASCII والرمز 178 وهـو يـشبه الرمز الذي تم إنشاؤه بالخيار {superscript:2} ولكن يظهر بمظهر أفـضل قليلاً، في المحور العمودي بالجانب الأيمن ASCII والرمز 186 يعرض رمز درجة الحرارة °C.

كيف نعرف أن الرمز 150 يمثل (–)، وأن الرمز 186 يمثل رمـز درجة الحرارة وهكذا، الشكل (16.3) يعرض جدولاً لرموز ASCII والتـي رُسِمتَ بأداة سهلة اسمها **asciiplot**، هذه الأداة ليست جزءًا من برنامج سـتاتا ولكنها لغة برمجة برنامج ستاتا وتم كتابة البرنامج الذي أنتج الجدول بالشكل (16.3) بواسطة عالِم الإحصاء الجغرافي نيكـولاس كـوكس Nicholas Cox، وللحصول على معلومات عن كيفية تنزيل وتحميل وكيفية عمل هـذه الأداة، قم بطباعة الأمر **findit asciiplot**، ويمكن إنشاء الجدول بطباعة الأمر **asciiplot** والذي يظهر في الشكل (16.3).

ASCII Code Character Map

use char(###) function to place symbols into graph text

First Digit(s) of ASCII Code	0	1	2	3	4	5	6	7	8	9
3				!	"	#	$	%	&	'
4	(	)	*	+	,	-	.	/	0	1
5	2	3	4	5	6	7	8	9	:	;
6	<	=	>	?	@	A	B	C	D	E
7	F	G	H	I	J	K	L	M	N	O
8	P	Q	R	S	T	U	V	W	X	Y
9	Z	[	\	]	^	_	`	a	b	c
10	d	e	f	g	h	i	j	k	l	m
11	n	o	p	q	r	s	t	u	v	w
12	x	y	z	{	|	}	~		€	
13	‚	ƒ	„	…	†	‡	ˆ	‰	Š	‹
14	Œ		Ž			‘	’	“	”	•
15	–	—	˜	™	š	›	œ		ž	Ÿ
16		¡	¢	£	¤	¥	¦	§	¨	©
17	ª	«	¬	-	®	¯	°	±	²	³
18	´	µ	¶	·	¸	¹	º	»	¼	½
19	¾	¿	À	Á	Â	Ã	Ä	Å	Æ	Ç
20	È	É	Ê	Ë	Ì	Í	Î	Ï	Ð	Ñ
21	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û
22	Ü	Ý	Þ	ß	à	á	â	ã	ä	å
23	æ	ç	è	é	ê	ë	ì	í	î	ï
24	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù
25	ú	û	ü	ý	þ	ÿ				

Last Digit of Code

الشكل (16.3)

الشكل (15.3) يوضح أبسط طـرق الـربط **connect** والتـي يمكـن استخدامها لربط نقاط البيانات بواسطة خط يربط بينهـا، الطـرق الأخـرى للربط تم عرضـها أدنـاه، الخطـوط الافتراضـية لـربط النقـاط هـي **connect(direct)** أو **connect(l)**، ولمزيد من المعلومات قم بطباعـة الأمـر **help connectstyle**.

connect()	الاختصار	الشرح
none	i	لا تربط النقاط
diect	l (حرف إل)	اربط النقاط بخطوط مستقيمة
ascending	L	ربط النقاط بطريقة مباشرة فقط في حالة أن $x[i+1]>x[i]$
stairstep	J	أفقي ثم عمودي
stepstair	(none)	عمودي ثم أفقي

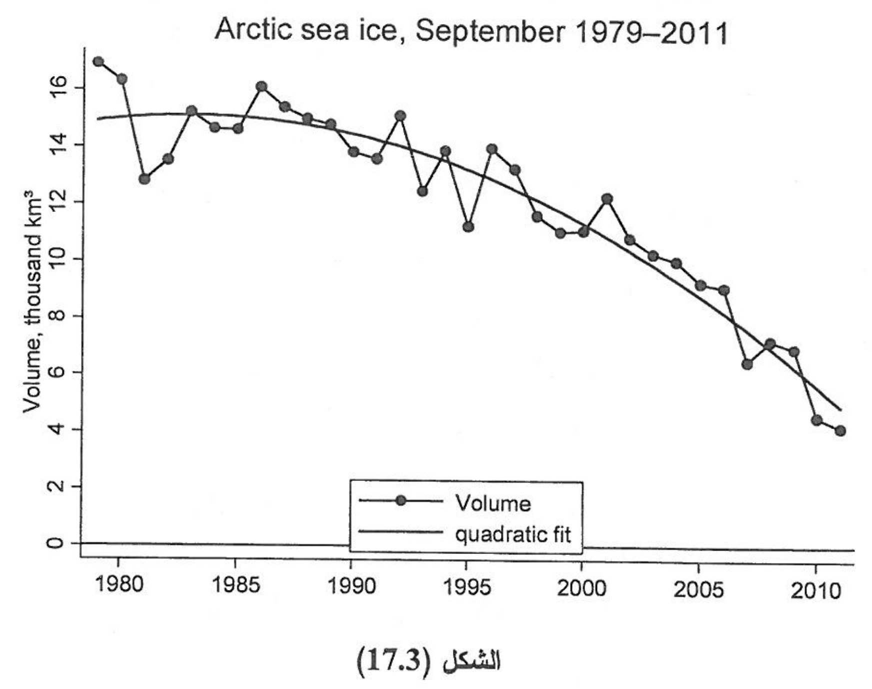
أنواع أخرى من الرسم البياني الثنائي : Other Twoway Plot Types

بالإضافة إلى الأنواع الرئيسة للرسم البياني الخطي والنقطي يمكن للأمر **graph twoway** رسم عدة أنواع أخرى. وفي هذا الجزء سوف يتم شرح عدة أنواع من الرسومات البيانية الثنائية. وللحصول على قائمة كاملة بهذه الأنواع قم بطباعة الأمر **help graph twoway**.

في وقت سابق عند إنشاء الشكل (10.3) والشكل (11.3) قمنا باستخدام الأمر **graph twoway lfit** وذلك لرسم خط الانحدار البسيط، هناك أمر مشابه للأمر السابق وهو **graph twoway qfit** الذي يقوم برسم منحنـى الارتبـاط المتعدد من الدرجة الثانية أو التربيعي، الشكل (17.3) يوضح استخدام متغير آخر من ملف البيانات *Arctic9.dta* وهو *volume*، وهذا المتغير تـم قياسـه بالكيلو متر المكعب، حيث إن الرمز 179 من ASCII يزودنا برمز المكعـب في كلمة km^3، ويجب الانتباه إلى أنه من الممكن استخدام **{superscript:3}**.

```
.graph twoway connect volume year,
 ylabel(0(2)16) yline(0)
 ytitle("Volume, thousand km`=char(179)'")
```

```
|| qfit volume year, lwidth(medthick)
|| , xlabel(1980(5)2010) xtitle("")
title("Arctic sea ice, September
1979`=char(150)'2011")
legend(row(2) ring(0) position(6)
label(1 "Volume") label(2 "quadratic fit")
order(1 2))
```

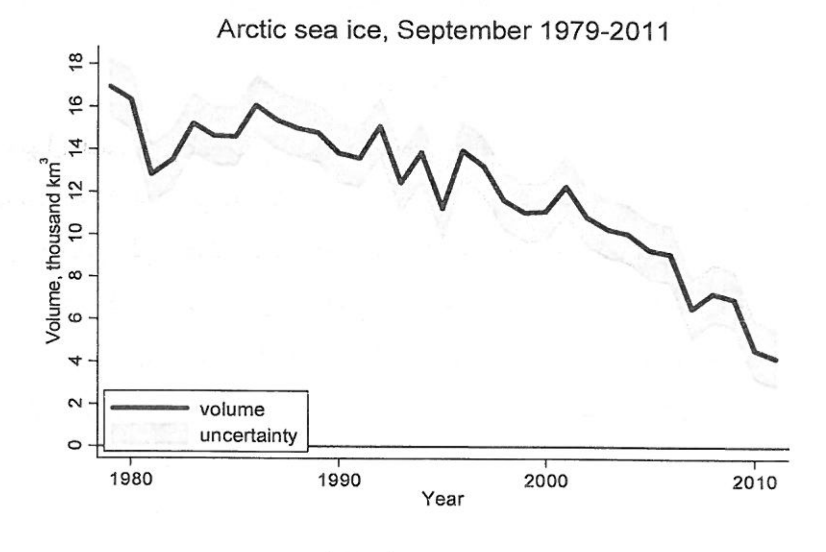


الشكل (17.3)

ميول المنحنى التربيعي نحو المحور الأفقي مقترباً من نقطة الصفر (yline(0)) يؤكد أن متوسط حجم الجليد في شهر سبتمبر انخفظ بشكل مستمر، الميول الخطي لن يعطي صورة واضحة عن الهبوط المتسارع وفي بعض الأحيان فإن هذا الميول غير واقعي، المنحنى في الشكل (17.3) يعرض ارتفاعاً بسيطاً جداً خلال السنوات الأولى وإذا توقعنا ما سوف يحدث في السنوات القادمة فإننا نتوقع أن حجم الجليد سوف ينخفض لأقل من الصفر؛ في الفصل (8) سوف نعود لهذه البيانات، ونأخذ في الاعتبار المنحنى بشكل أكثر دقة.

الشكل (18.3) ينظر للمتغير *volume* من زاوية مختلفة. ففي هذه المرة يستخدم الرسم البياني لمنطقة المدى **graph twoway rarea** ليوضح بأن 95% من حدود عدم التأكد زائد أو ناقص 1.35 ألف كيلـومتر مكعـب (دراسـة Schweiger et al 2011) منطقة المدى أو الأمر **rarea** يبحث عن متغيرين في المحور العمودي *y* . وفي هذا المثال هذان المتغيران هما المتغير *volumehi* والمتغير *volumelo* واللذان يحددان الحدود العليا والسفلى لمنطقة معينة ليـتم تظليلها، الرسم البياني الخطي للمتغير *volume* تم رسـمه بخـط عـريض **(lwidth(thick))** وتم وضع هذا الشكل فوق الرسم البياني الذي يحدد منطقـة المدى **rarea** حتى يمكن مشاهدة المنطقة المظللة.

```
.graph twoway rarea volumehi volumelo year,
 color(gs13)
 || line volume year, lwidth(thick)
 lcolor(green)
 ||, ylabel(0(2)18) yline(0)
 ytitle("Volume, thousand km'=char(179)'")
 title("Arctic sea ice, September
 1979'=char(150)'2011")
 legend(row(2) ring(0) position(7)
 label(1 "uncertainty") label(2 "volume") order
 (2 1))
```

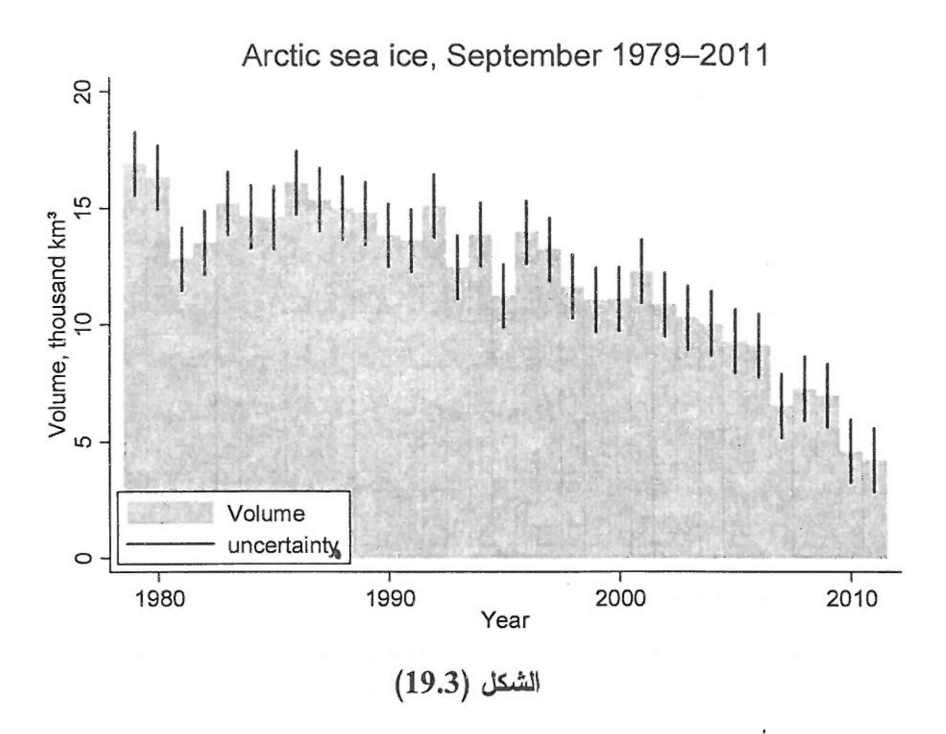


الشكل (18.3)

الأمر **graph** الذي تم استخدامه في رسم الشكل (18.3) يحدد ألوان خط الرسم البياني ولون منطقة المدى، أما الخيار **color(gs13)** فهو يعطي منطقة المدى اللون الرمادي 13، وهو أقرب للون الأبيض **(gs16)** منه للون الأسود **(gs0)**، الرسم البياني الخطي تم وضعه فوق هذا الرسم، والخط ذو لون أخضر؛ التحكم في الألوان (بدلاً من استخدام اللون الافتراضي) يقوم بوظائف محدودة في حالة النشر باللونين الأبيض والأسود ولكنه فعّال جداً في حالة استخدام عدة ألوان، ولمشاهدة قائمة بالألوان المتوافرة للرسومات البيانية في ستاتا قم بطباعة الأمر **help colorstyle**.

الشكل (19.3) يضع شكلين من النوع الثنائي لإنشاء صورة مختلفة من نفس المعلومات الخاصة بحجم الجليد في القطب الشمالي. في هذا الشكل البياني المتغير *volume* تم تمثيله بأعمدة **graph twoway bar** بينما مدى عدم التأكد ظهر كرؤوس مدببة **(rspike)** بدلاً من نطاق ملون **(rarea)** والذي سبق أن شاهدناه في الشكل (18.3)، الأعمدة تم تلوينها بلون رمادي خفيف **(color(gs10))**، ويجب ملاحظة أن الشكل (18.3) وضع خط المتغير *volume* فوق منطقة مدى عدم التأكد، بينما وضع الشكل (19.3) الرؤوس المدببة التي تمثل مدى عدم التأكد في أعلى الأعمدة البيانية التي تمثل المتغير *volume*، الترتيب في الشكلين مطلوب، حتى نستطيع مشاهدة الشكل بوضوح، ويمكنك اختبار ذلك مع أوامر مشابهة لمشاهدة كيفية عمل تلك الأوامر.

```
.graph twoway bar volume year, color(gs10)
|| rspike volumehi volumelo year,
 lwidth(medthick)
|| , ytitle("Volume, thousand km`=char(179)'")
title("Arctic sea ice, September
 1979`=char(150)'2011")
legend(row(2) ring(0) position(7)
label(1 "Volume") label(2 "uncertainty"))
```



الشكل (19.3)

لا يجب الخلط بين الرسم البياني الثنائي المضلع **graph twoway bar** والمضلع التكراري، حيث يوجد اختلاف بين الأمرين، فالأمر **twoway bar** يقوم ببساطة بإنشاء رسم بياني للمتغير y مع المتغير x مفترضاً أن المتغيرين تم قياسهما بوحدات قياس مختلفة، ولكنهما يشتركان في خاصيات متعددة مثل توصيف المحورين الأفقي والعمودي، واحتمالية وضع الرسومات فوق بعضها.

ستاتا يوفر أكثر من 40 نوعاً مختلفاً من أنواع الرسم البياني الثنائي **graph twoway** لا يسع المجال لذكرها هنا. في الفصول القادمة، سوف نشرح أنواعاً أخرى من هذه الرسومات، وللحصول على قائمة كاملة بالأوامر الخاصة بهذه الرسومات وتركيباتها، قم بطباعة الأمر **help graph twoway**.

الرسومات البيانية العمودية والدائرية : Bar Charts and Pie Charts

الأمر **graph bar** يختلف عن الأمر **graph twoway bar**، حيث إن الأول يعرض العلاقات التي تتضمن نوعاً واحدًا أو أكثر من المتغيرات المصنفة

مثل هذه الأشكال البيانية تبرهن بشكل خاص على أهمية بيانات الدراسات الاستقصائية، كما سوف يتم عرضه في الفصل (4). هذا الجزء يعتبر كمقدمة لهذا الأمر مع أمثلة باستخدام متغيرات من ملف بيانات *Nations_2.dta*

```
.use C:\data\Nations2.dta, clear
.describe region gdp pop
```

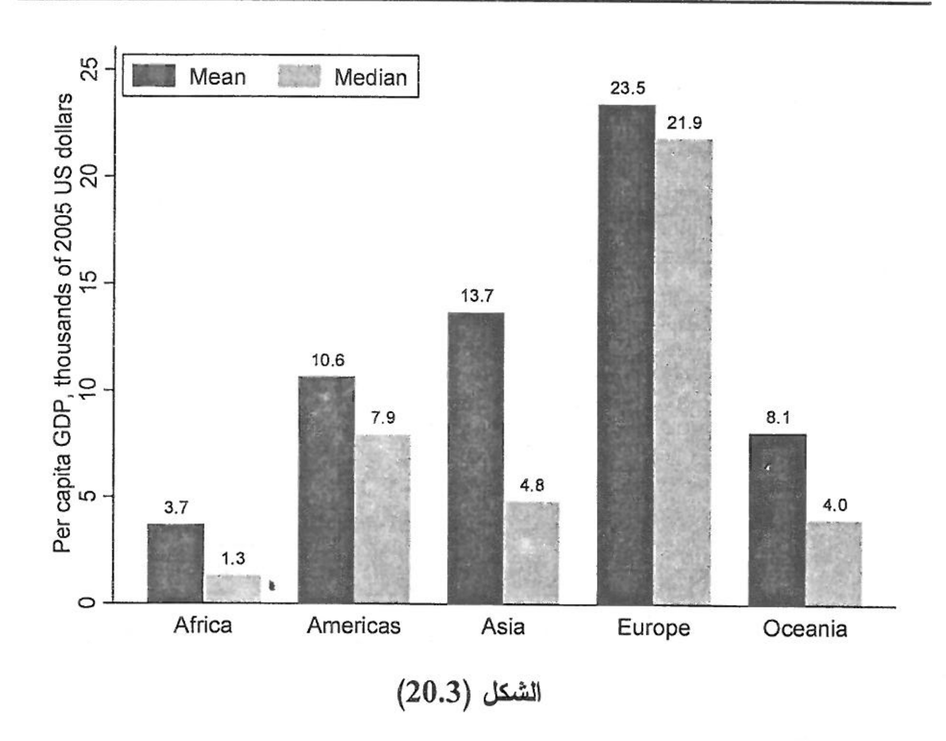
variable name	storage type	display format	value label	variable label
region	byte	%8.0g	region	Region
gdp	float	%9.0g		Gross domestic product per cap 2005$, 2006/2009
pop	float	%9.0g		Population 2005/2010

المتغير *gdp* يمثل الناتج المحلي الإجمالي للفرد. وقيمه تتراوح بين 279.8 إلى 74.906 دولار للفرد، استخدام 5 أرقام كتوصيف للمحور الأفقي أو العمودي أمر غير عملي، لذلك سوف نبدأ بإنشاء متغير جديد باسم *gdp1000* يوضح الناتج المحلي الإجمالي بآلاف الدولارات، الشكل (20.3) يعرض المتوسط والمنوال للمتغير *gdp1000* للقارات *region* باستخدام الرسم البياني العمودي **graph bar**.

```
.generate gdp1000 = gdp/1000
.summarize gdp gdp1000
```

Variable	Obs	Mean	Std. Dev.	Min	Max
gdp	179	12118.74	13942.34	279.8	74906
gdp1000	179	12.11874	13.94234	.2798	74.906

```
.graph bar (mean) gdp1000 (median) gdp1000,
over(region)
ytitle("Per capita GDP, thousands of 2005 US
dollars")
blabel(bar, format(%3.1f))
bar(1, color(blue)) bar(2, color(orange))
legend(ring(0) position(11) col(2) label(1
"Mean")
label(2 "Median") symxsize(*.5))
```



الشكل (20.3)

الشكل (20.3) يتضمن توصيفات تظهر في أعلى الأعمدة بالرسم البياني، ويعرض تنسيقات للأرقام **format(%3.1f)**، والذي يعني تنسيقات ثابتة مع ثلاثة أرقام، أحدها على يمين الفاصلة العشرية، والأعمدة يمكنها ليس فقط عرض المتوسطات أو المنوال وإنما أيضاً عرض عدد من الحسابات الإحصائية الأخرى مثل النسب المئوية، وأعلى قيمة، وأقل قيمة والأعداد. كما يمكنها أيضاً أن تعرض هذه الإحصائيات لأكثر من متغير واحد إذا تشابهت وحدة قياس هذه المتغيرات.

مربع شرح الرسم في الشكل (20.3) يظهر في داخل الرسم نفسه في ناحية البيانات بموقع الساعة 11: **ring(0) position(11)** وبه عمودان متوازيان جنباً إلى جنب، أما الخيار **symxsize(*.5)** يجعل حجم الرموز في المحور الأفقي في مربع شرح الرسم تظهر بنصف حجمها الافتراضي، وذلك لتوفير مساحة في الرسم. كما يجعل الرموز أكثر تشابهاً مع حجم الأعمدة نفسها.

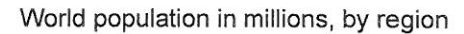
الرسم البياني العمودي **graph bar** في المثال أعلاه، يحدد اللون الأزرق للعمود 1، واللون البرتقالي للعمود 2. اللونـــان الأزرق والبرتقـــالي قـــد لا يظهران في هذه الصفحة بسبب الطباعة باللونين الأبيض والأســود، ولكـــن الاختلاف بينهما واضح، وكما أشار نيكولاس كـــوكس Nicholas Cox فـــإن اللونين الأزرق والبرتقالي من الألوان التي تلفت الانتباه، لأنها تُظهِر اختلافاً مرئياً واضحاً للقارئ عن أغلب الألوان الأخرى، فهي مـــثلاً تختلـــف عـــن الأحمر والأخضر. المحللون ربما يأخذون هذه الاختلافات في الاعتبار عند تصميم الرسومات البيانية خصوصاً عندما تكون التفرقة بين الألوان أمرًا في غاية الأهمية.

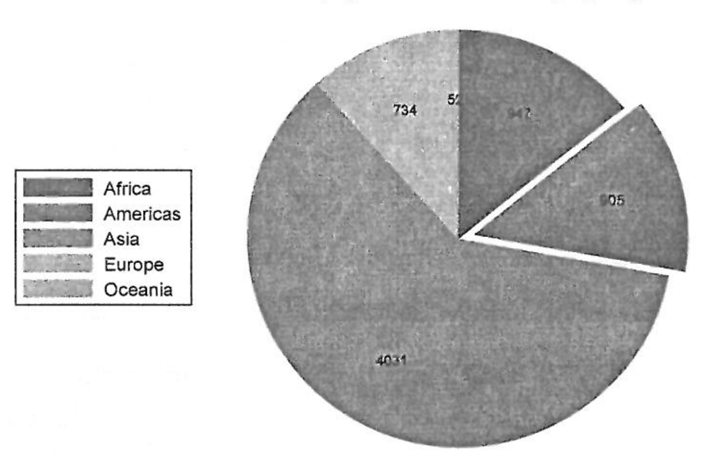
الرسومات البيانية العمودية يمكنها تـــوفير شـــرح واضـــح للعلاقـــات المتداخلة بين العديد من الفئات وعند وجود متغيرين أو أكثر. ومـــن ناحيـــة أخرى، فإن الرسم البياني الدائري نادراً ما يوضّح التحليل، ولكنه شائع فـــي بعض المحاضرات العامة؛ الشكل (21.3) يشرح أمر إنشاء الرسم البيـــاني الدائري **graph pie** موضحاً عدد السكان لكل قارة، فالمتغير *pop* يمتد من أقل مـــن 10,000 إلـــى 1.32 مليـــار نـــسمة (1.32e يعنـــي $1.32 \times 10^9 =$ 1,320,000,000)، ولإنشاء رسم بياني دائري سهل القـــراءة، فإنـــه مـــن الإفضل إنشاء متغير جديد باسم *popmil* أو عدد السكان بالمليون نسمة.

```
.gen popmil=pop/1000000
.summarize pop popmil
```

Variable	Obs	Mean	Std. Dev.	Min	Max
pop	194	3.44e+07	1.31e+08	9767	1.32e+09
popmil	194	34.37752	131.4004	.009767	1324.696

```
.graph pie popmil, over(region) pie(2, explode)
plabel(_all sum, format(%4.0f))
title("World population in millions, by
region")
legend(col(1) position(9))
```





الشكل (21.3)

الخيار **pie(2,explode)** يقوم بإبراز الشريحة الثانية (Americans) خارجاً للتأكيد على أهميتها، أما الخيار **plabel(_all sum, format(%4.0f)** يحدد توصيفات لكل الفئات بالشكل الدائري معطياً مجموع المتغير *popmil* (مجموع عدد السكان بالمليون نسمة) لكل منطقة، أما توصيفات الرسم البياني الدائري تم تنسيقها بالأمر **format(%4.0f)** والذي يعني تنسيقاً رقمياً ثابت مع أربعة أرقام بدون وجود أي أرقام بعد الفاصلة.

لمشاهدة كل الخيارات المتوافرة مع الرسم البياني الدائري، قم بطباعة الأمر **help graph pie** حيث إن الخيارات تتضمن الطرق المختلفة لتنظيم البيانات، أحد الخيارات المثيرة هو **by()** والذي يقوم بإنشاء صورة تحتوي على رسومات بيانية متعددة يمكن مشاهدتها والمقارنة بينها.

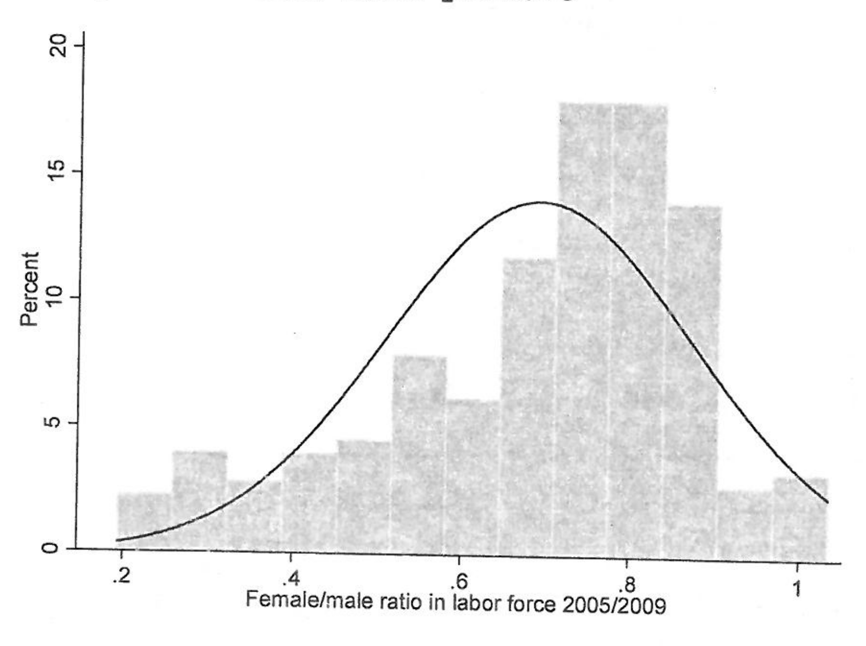
الرسم البياني للربيعات والرسم البياني التناظري :
Symmetry and Quantile Plots

رسم الصندوق، والرسم البياني العمودي، والمدرج التكراري، تلخص توزيع قياس المتغيرات، ولكنها لا توضح نقاطاً مهمة بالبيانات، فهي عبارة عن نظرة عامة على البيانات ككل. ومن ناحية أخرى، فإن الرسم البياني للربيعات، والرسم البياني التناظري، يتضمنان نقاطاً لكل مشاهدة. فالقارئ

يحتاج إلى مجهود أكثر لقراءتهما، لأنهما يعرضان تفاصيل أكثر من تلك التفاصيل التي يعرضها الرسم البياني العمودي، والمدرج التكراري.

المدرج التكراري لنسبة الإناث إلى الذكور في القوى العاملة لـ 177 دولة الموجودة في ملف البيانات *Nations2.dta*– والتي تظهر في الشكل (22.3) تم تركيبها فوق منحنى التوزيع الطبيعي (توزيع جاوس) الذي يشير إلى أن المتغير *femlab* (الذي يمثل نسبة الإناث في القوى العاملة) له ذيل أكثر من الطبيعي إلى جهة اليسار (الدول التي بها عدد إناث أقل في القوى العاملة) وذيل أقل من الطبيعي جهة اليمين وهذا يعني وجود التواء سالب.

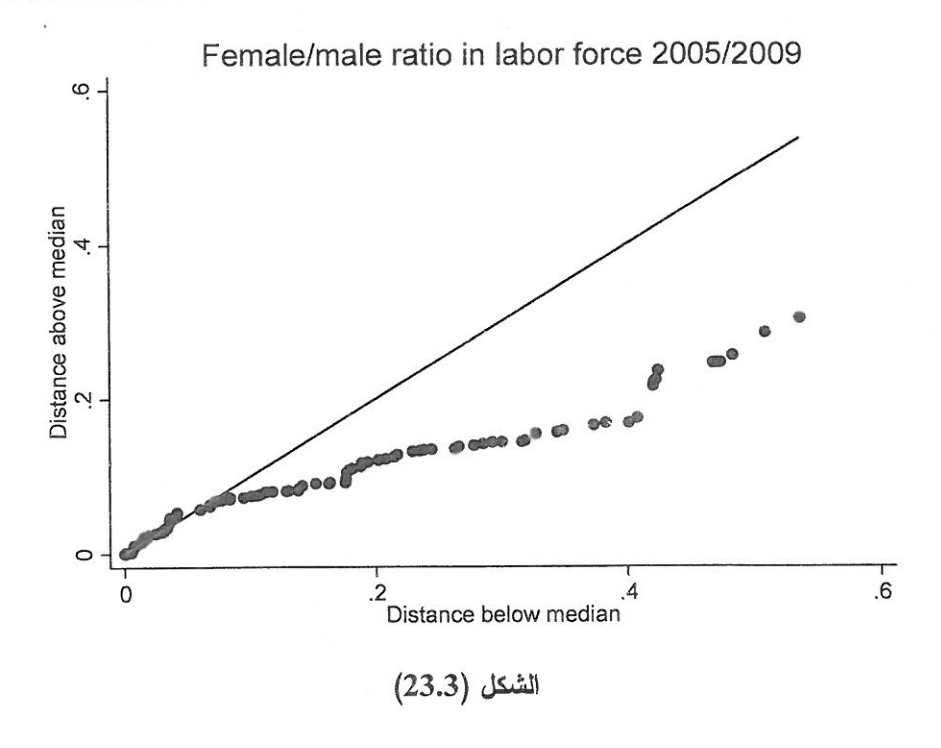
```
.histogram femlab, norm percent
```



الشكل (22.3)

الشكل (23.3) يعرض هذا التوزيع في رسم بياني تناظري موضحاً المسافة بين المشاهدة *i*th التي تقع أعلى من الوسيط (العمودي) والمشاهدة *i*th التي تقع تحت الوسيط. كل النقاط في الرسم كان سيتم وضعها على الخط المحوري لو كان التوزيع تناظرياً. ولكننا نرى وجود مسافة تحت الوسيط، وهذه المسافة تزداد بمقدار أكبر مشيرةً إلى التواءٍ سالب.

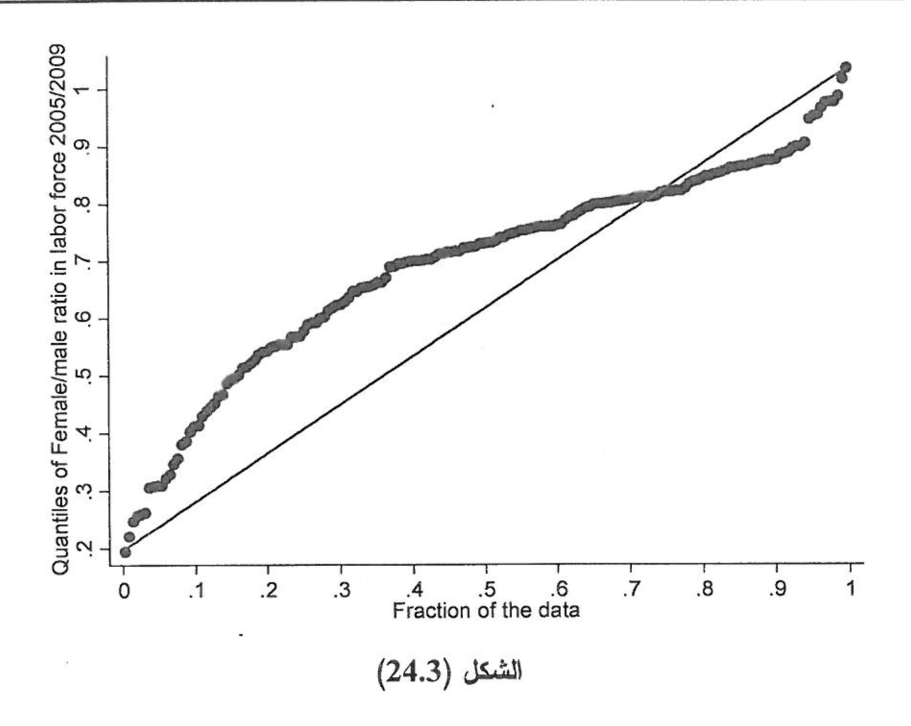
```
.symplot femlab
```



الشكل (23.3)

الربيعات هي القيم التي تظهر في أسفل مستوى معين من البيانات. فمثلاً الربيع 0.3، هو عبارة عن القيم الأكبر من 30% من البيانات (تشبه الـ 30%)، وإذا قمنا بترتيب عدد المشاهدات n تصاعدياً، فإن قيمة المشاهدة ith تمثل الربيع $(i\text{-}5)/n$، والرسم البياني للربيعات يقوم آلياً بحساب جزء المشاهدات الذي تقع تحت كل قيمة في البيانات، ويعرض النتائج برسم بياني في الشكل (24.3)، والرسم البياني للربيعات يعتبر مرجعاً لأي شخص ليس لديه البيانات الأصلية، ومن خلال التوصيفات الموجودة بالرسم البياني، يمكننا تقدير بعض الحسابات الإحصائية مثل الوسيط (الربيع 0.5) أو العشيرات (الربيعات 0.1، 0.2، 0.3 وهكذا) كما يمكننا أيضاً قراءة الرسم البياني للربيعات لتقدير جزء المشاهدات التي تقع تحت قيمة محددة معطاة.

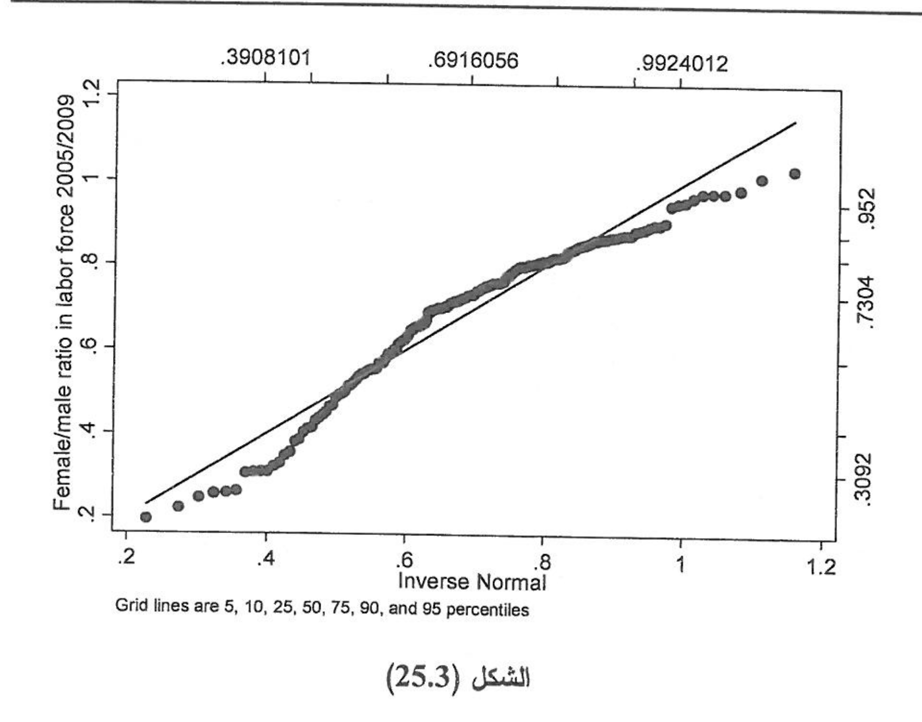
```
.quantile femlab, xlabel(0(.1)1,grid)
 ylabel(.2(.1)1,grid)
```



الشكل (24.3)

الرسومات البيانية للربيعات العادية تسمى أيضاً الرسومات البيانية للاحتمال الطبيعي، التي تقارن بين توزيع متغير ما مع ربيعات توزيع طبيعي نظري له نفس المتوسط والانحراف المعياري، هذه الرسومات تسمح لنا بإجراء فحص نظري للانحراف عن التوزيع الطبيعي لكل جزء من التوزيع، والذي يمكن أن يساعد كدليل في اتخاذ القرارات التي تتعلق بالافتراضات الطبيعية والجهود التي تحاول جعل التحول طبيعياً أكثر. الرسم البياني للربيعات العادية بالشكل (25.3) للمتغير *femlab* يؤكد لنا الالتواء السالب والذي سبق ملاحظته سابقاً، والخيار **grid** يقوم بإنشاء خطوط شبكة لـ 0.05، 0.10، 0.25 (الربيع الأول) و0.50 (الوسيط) و0.75 (الربيع الثالث) والربيعات 0.90 و0.95 للتوزيعات، قيم الربيعات 0.05 و0.50 و0.95 تم عرضها في أعلى ويمين المحاور.

```
.qnorm femlab, grid
```



الشكل (25.3)

إضافة نص للرسومات البيانية : Adding Text to Graphs

يمكن إضافة عناوين وشروحات وملاحظات للرسم البياني لجعله أكثر وضوحاً، ففي الشكل الافتراضي العناوين الرئيسة والعناوين الفرعية تظهر فوق منطقة البيانات (والتي ربما توثق مصدر البيانات مثلاً)، ومربع شرح الرسم يظهر تحت منطقة الرسم، انظر الشكل 7.3) على سبيل المثال، والذي يستخدم عنواناً وشرحاً وملاحظة. هذه الخيارات الافتراضية يمكن تجاوزها. فمثلاً للحصول على معلومات أكثر عن خيارات العناوين، قم بطباعة الأمر **help title options**، وعن خيارات تغيير المحتويات مثل حجم الخط ولونه قم بطباعة الأمر **help textbox options**.

كما أنه من الممكن إضافة مربعات نصيّة في مواقع محددة داخل منطقة الرسم نفسه. فعلى سبيل المثال، ربما نريد إضافة ملاحظة في رسم بياني لسلسلة زمنية مثل تواريخ ذوبان الجليد في بحيرة وينبيسوكي Lake Winnipesaukee كما تم الإشارة إليها في الشكل (26.3).

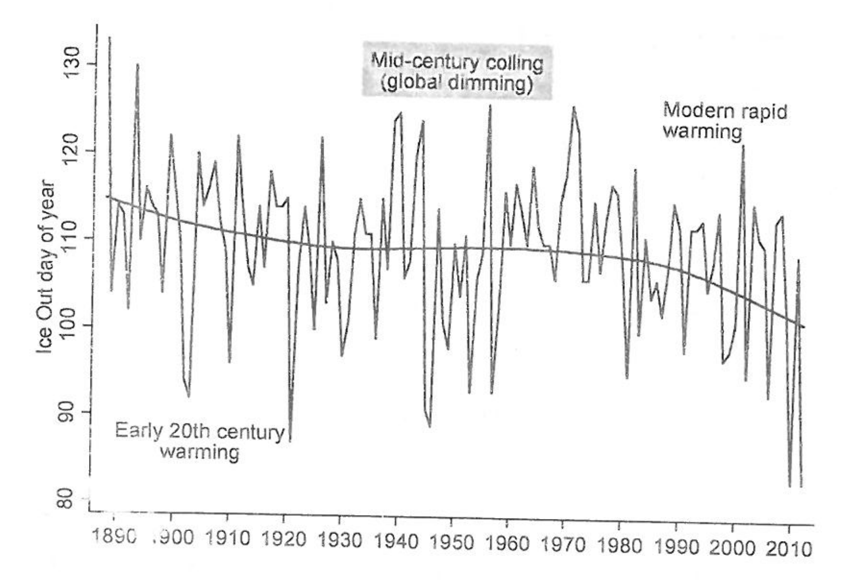
```
.use C:\data\lakesunwin.dta, clear
.describe
```

```
Contains data from c:\data\lakesunwin.dta
  obs:           144                     Sunapee & Winnipesaukee ice out 1869-2012
 vars:             5                     4 Jul 2012 11:21
 size:         1,728
```

variable name	storage type	display format	value label	variable label
year	int	%ty		Year
sunedate	float	%tdCYmd		Date Lake Sunapee Ice-Out
sunout	int	%9.0g		Lake Sunapee Ice-Out day
winedate	int	%tdCYmd		Lake Winnipesaukee Ice-Out
winout	int	%9.0g		Lake Winnipesaukee Ice-Out day

```
Sorted by:  year
```

```
.graph twoway line winoutyear
|| lowess winout year, lwidth(medthick)
|| if year>=1887 , xlabel(1890(10)2010)
legend(off) xtitle("")
ytitle("Ice Out day of year")
text(87 1905 "Early 20th century" "warming")
text(130 1950 "Mid-century cooling" "(global
dimming)",
box margin(small) bcolor(gs11))
text(125 1998 "Modern rapid" "warming",
justification(left))
```



الشكل (26.3)

الرسم البياني المبسط في الشكل (26.3) من نوع **twoway line** يمثل أيام ذوبان الجليد في بحيرة Winnipesaukee للفترة من سنة 1887 إلى 2012، تم وضع شكل بياني ثان يعرض منحنى انحدار **lowess** تم رسمه بخط متوسط العرض (**lwidth(medthick)**) ليكون أكثر وضوحاً، وانحدار القطع التمهيدي – سوف يتم شرحه في الفصل (8) – يعتبر مدخلاً مرناً لمعرفة تجانس البيانات، وله عدة مزايا تجعله أفضل من طرق المتوسط المرجح، أما منحنى القطع التمهيدي في الشكل (26.3) يوضح اتجاهات عدة عقود التي تكمن وراء التغيرات التي حدثت من سنة لأخرى.

الاتجاهات خلال عقود في وضع البحيرة تتبع بشكل عام نفس نمط درجات الحرارة العالمية، وبيانات ولاية نيوهامبشير خلال نفس الفترة التي تم شرحها في دراسة (Hamilton et al. 2010a)، أما تواريخ ذوبان الجليد، فقد شهدت انخفاضاً عاماً خلال فترة ارتفاع درجات الحرارة في نهاية القرن 19 وبداية القرن 20، وكان هناك انخفاض بسيط في درجات الحرارة خلال الفترة من الأربعينيات إلى بداية السبعينيات، مما أدى إلى تغير في تواريخ ذوبان الجليد. وعالمياً فإن هذه الفترة تمثل المرحلة التي كانت تصل فيها أشعة الشمس إلى سطح الجليد بسبب التلوث الصناعي حسب دراسة (Wild et al. 2007)، وخلال فترة زيادة درجة الحرارة والتي كانت ما بعد منتصف السبعينيات كانت تواريخ ذوبان الجليد في بحيرة Winnepesaukee قد نقصت بدرجة كبيرة، حيث إن أحدث بيانات مسجلة لتواريخ ذوبان الجليد كانت في سنة 2010 وسنة 2012.

ثلاثة مربعات شرح نصية تم وضعها في منطقة البيانات بالشكل (26.3) توضح التغيرات المناخية. حيث يحتوي كل مربع على سطرين، تم الفصل بينهما بواسطة الخيار **text()** مع علامتي اقتباس، فالخيارات **text()** تحدد مربعات شرح الرسم البياني والتي عادة ما تحتاج إلى اختبار تحديد موقعها بالضبط، فمربع شرح الرسم الأول تم إنشاؤه بالأمر:

```
text(87 1905 "Early 20th century" "warming")
```

حيث يقوم بوضع مربع الشرح في الموقع $y=87$ و$x=1905$ مع استخدام الخيارات الافتراضية، أما مربع الشرح الثاني فيتم وضعه عند $y=130$ و$x=1950$، حيث يُعرض مُحاطاً بحدود يمكن مشاهدتها بوضوح مع هامش صغير **small** حول النص، أما ألوان الحدود والخلفية فيتم تحديدها بلون رمادي **gs11**

```
text(130 1950 "Mid-century colling" "(global
dimming)",
box margin(small) bcolor(gs11))
```

مربع الشرح الثالث يحتوي على نص تم محاذاته لجهة اليسار:

```
text(125 1998 "Modern rapid" "warming",
justification(left))
```

وللحصول على مزيد من المعلومات حول التحكم في مربعات الشرح النصية يمكنك طباعة الأمر **help textbox options** والأمر **help colorstyle**

الرسم البياني مع ملفات التنفيذ Do-Files :

Graphing with Do-Files

الأشكال البيانية المعقدة مثل الشكل (26.3) تتطلب أسطر أوامر **graph** طويلة (بالرغم من أن ستاتا يعرض الأمر بالكامل في سطر واحد)، ملفات التنفيذ Do-Files – التي تم شرحها في الفصل (2) – تساعد في كتابة الأوامر ذات الأسطر المتعددة، كما أنها تسهل حفظ الأمر، وإعادة استخدامه مستقبلاً في حالة الحاجة لتعديل الرسم البياني أو رسمه من جديد.

الأوامر التالية تم طباعتها في محرر Do-File، وحفظها في ملف باسم *fig03_26.do* ، حيث إنها تمثل ملفاً جديدًا لرسم الشكل (26.3) من جديد.

```
.use C:\data\lakesunwin.dta, clear
.graph twoway line winout year ///
|| lowess winout year, lwidth(medthick) ///
|| if year>=1887 , xlabel(1890(10)2010) legend(off)
xtitle("") ///
ytitle("Ice Out day of year") ///
text(87 1905 "Early 20th century" "warming")
///
text(130 1950 "Mid-century cooling" "(global
dimming)", ///
```

```
box margin(small) bcolor(gs11)) ///
text(125 1998 "Modern rapid" "warming", ///
justification(left))
.graph save Graph C:\graphs\fig03_26.gph,
replace
.graph export C:\graphs\fig03_26.png, as(png)
replace
.graph export C:\graphs\fig03_26.eps, as(eps)
replace
```

عندما ينتهي السطر بعلامات /// في الملف التنفيذي do-file فإن ذلك يعني أن الأمر لم يكتمل بعد ومازال مستمرًا في السطر التالي. والأمر يتم تنفيذه فقط عندما لا تحتوي نهاية السطر على ///، أماالطريقة الأخرى لكتابة الأوامر ذات الأسطر المتعددة فتتم باستخدام الأمر **#delimit;** والذي يقوم بتحديد نهاية الأمر باستخدام فاصلة منقوطة بدلاً من النهاية الافتراضية وهي ضغط مفتاح Enter أو استخدام **(#delimit cr)** التي سبق شرحها في الفصل (2).

الأمر **graph save Graph** يقوم بحفظ الصورة أو الرسم البياني (والتي يتم إعطاؤها اسماً افتراضياً مؤقتاً "Graph" والتي تظهر في نافذة الرسم Graph window) بتنسيق ستاتا الذي ينتهي بالامتداد gph. كما يمكننا تحديد اسم مؤقت لأي رسم بياني وذلك بإضافة خيارات أخرى للأمر **graph** مثل خيار **name(*newgraph*)** أو خيار **name(*fig03_26*)**، استخدام مثل هذه الأسماء المؤقتة للرسومات البيانية يُعتبر أمرًا مهماً جداً عندما يكون لدينا مجموعة من الرسومات البيانية المعروضة، ونريد حفظ أو طباعة أحدها، كما أن إعطاء اسم مؤقت للرسم البياني الذي قمنا بإنشائه لا يعني حفظه في القرص، وليس من الضروري تطابق أسماء الملفات المؤقتة والمحفوظة، وللحصول على معلومات أكثر عن خيارات الحفظ قم بطباعة الأمر **help name option**.

الأمر **graph export** الموجود بالملف do-file يقوم بإنشاء إصدار ثان وثالث لنفس الرسم البياني بتنسيقات مختلفة. فالملفات ذات التنسيق (.png) تعتبر صورًا نقطية ولها دقة ثابتة وقابلة للاستخدام مع برامج أخرى، ويمكن مشاركتها من خلال صفحات الويب ومايكروسوف بوربوينت وغيرها من

التطبيقات، أما الملفات ذات التنسيق (eps.) فهي ذات جودة عالية وتنسيقاتها مفضلة في النشر، ولمعرفة المزيد عن الخيارات الأخرى، يمكنك طباعة الأمر **help graph export**.

عند حفظ أوامر إنشاء وحفظ الرسومات البيانية في ملف تنفيذي باسم *fig03_26.do* يمكن تطبيق أوامر هذا الملف بطباعة الأمر التالي:

```
.do fig03_26
```

الأمر أعلاه سوف يقوم بتنفيذ كل الأوامر التي يحتويها الملف التنفيذي، حيث سيقوم بإنشاء الرسم البياني وحفظه في ثلاثة تنسيقات مختلفة، وسوف يتم الحفظ فوق الملفات الموجود مسبقاً.

استعادة ودمج الرسومات البيانية :

Retrieving and Combining Graphs

أي رسم بياني مخزن ببرنامج ستاتا بتنسيق gph يمكن استعادته في ذاكرة الجهاز باستخدام الأمر **graph use**. فعلى سبيل المثال، يمكننا استعادة الشكل (26.3) بطباعة الأمر:

```
.graph use fig03_26.gph
```

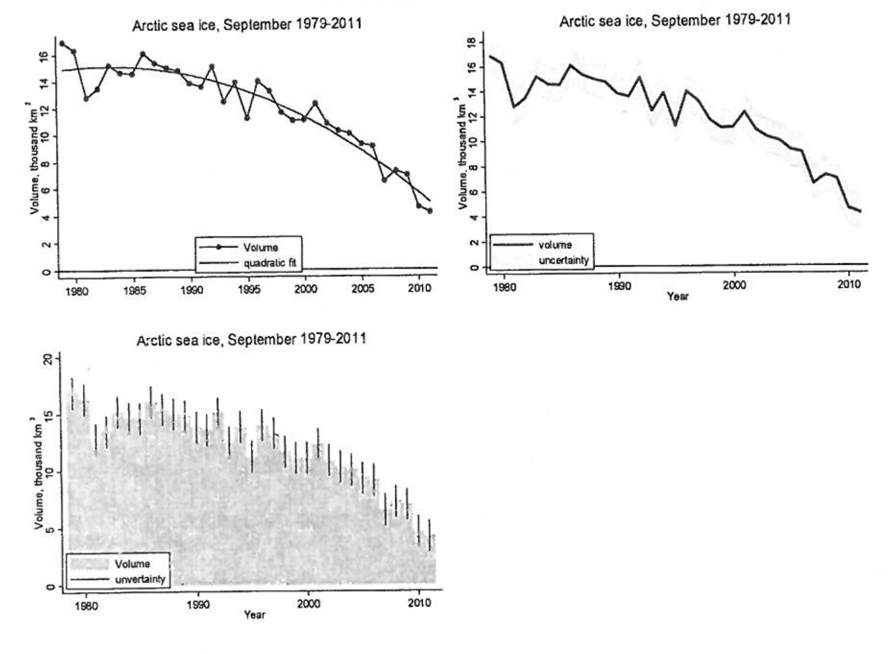
عندما يظهر الرسم البياني في ذاكرة الحاسب، فإنه يُعرض على الشاشة ويمكن طباعته أو حفظه مرة أخرى باسم مختلف أو بامتداد مختلف، فمثلاً إذا كان لدينا رسم بياني تم حفظه سابقاً بامتداد gph يمكننا إعادة حفظه باستخدام امتدادات مختلفة مثل eps, .png, .emf. كما أنه من الممكن تغيير لون الرسم باستخدام قوائم ستاتا أو باستخدام الأمر **graph use**. فمثلاً الملف *fig03_26.gph* الذي تم حفظه سابقاً كانت ألوانه بتنسيق s2color، ولكن يمكننا أن نرى كيف يظهر الشكل البياني باستخدام تنسيق الألوان s2monochrome (وهي بديل للخطوط المتقطعة بعدة ألوان) وذلك بطباعة الأمر:

```
.graph use fig03_45.gph, scheme(s2mono)
```

الرسومات البيانية المخزنة على القرص الصلب يمكن دمجها باستخدام الأمر **graph combine** بحيث يمكن عرض عدة رسومات بيانية في نفس الإطار.

ملف التنفيذ do-file أدناه (تم حفظه باسم *fig03_27.do* ثم أُعيد تشغيله بطباعة الأمر **do fig03_27**) يدمج الشكل (17.3) والشكل (18.3) والشكل (19.3) والتي تم إنشاؤها في هذا الفصل، وجود /// يشير إلى استمرارية الأمر في السطر التالي. والشكل النهائي الذي يظهر الرسومات المدمجة تم حفظه وإعطاؤه اسم الشكل (27.3).

```
.graph combine ///
C:\A_books\SwS_12\graphs\fig03_17.gph ///
C:\A_books\SwS_12\graphs\fig03_18.gph ///
C:\A_books\SwS_12\graphs\fig03_19.gph ///,
rows(2) altshrink ///
title("Combining Figures 3.17-19",
size(medium))
.graph save Graph C:\graphs\fig03_27.gph,
replace
.graph export C:\graphs\fig03_27.emf, as(emf)
replace
```



الشكل (27.3)

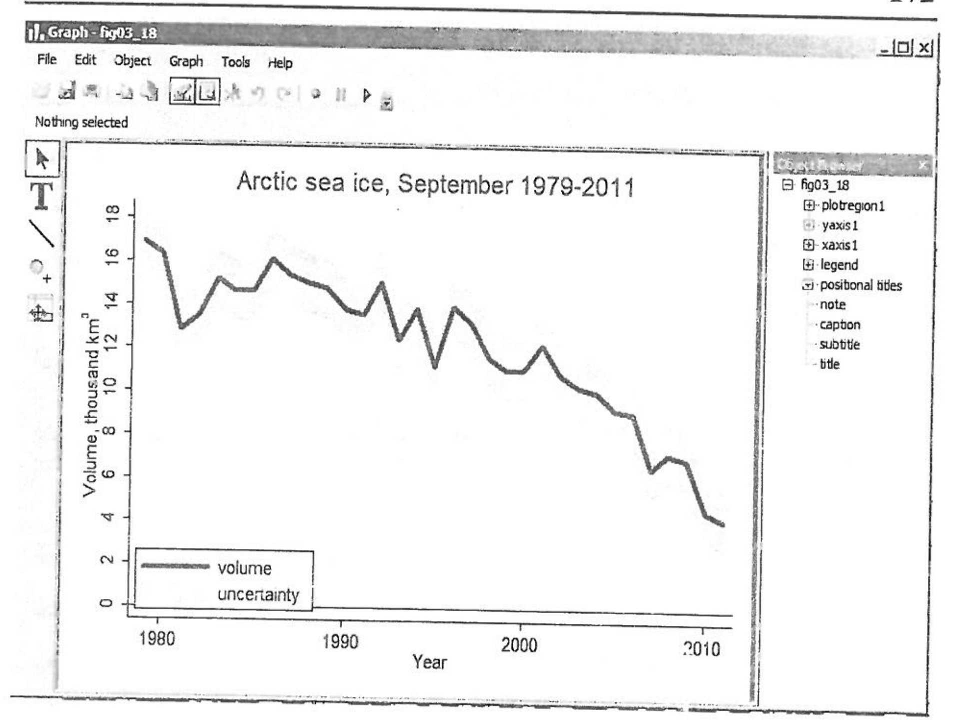
الخيار rows(2): يحدد أن الشكل (27.3) يجب أن يتم ترتيبه في شكل بياني فرعي في صفين اثنين، كما أنه من الممكن تحديد الخيار col(2).altshrink والذي يحدد تتابع النصوص في كل شكل بياني صغير في الشكل البياني (27.3) ويجب ملاحظة أنه بالإمكان وضع عنوان عام (أو ملاحظة أو شرح أو عناوين للمحاور.. الخ) للشكل البياني ككل، ولكن لا يمكننا إجراء تغيير جوهري في محتويات الأشكال الفرعية الصغيرة بالرسم البياني.

محرر الرسم البياني : Graph Editor

محرر الرسم البياني Graph Editor يتيح لنا تعديل شكل الرسم البياني الموجود حالياً في الذاكرة سواءً كان هذا الشكل قد تم إنشاؤه الآن أو سبق حفظه من قبل وتم استعادته باستخدام الأمر **graph use**، إنه من الأسهل أن تعرف معلومات حول خاصية تحرير الرسم بالتدريب عليها واختبارها الآن بدلاً من الاطلاع عليها نظرياً فقط، فمثلاً إذا كنا نريد القيام ببعض التغييرات في الشكل (18.3) نبدأ أولاً باستعادة هذا الشكل باستخدام الأمر:

```
.graph use C:\graphs\fig03_18.gph
```

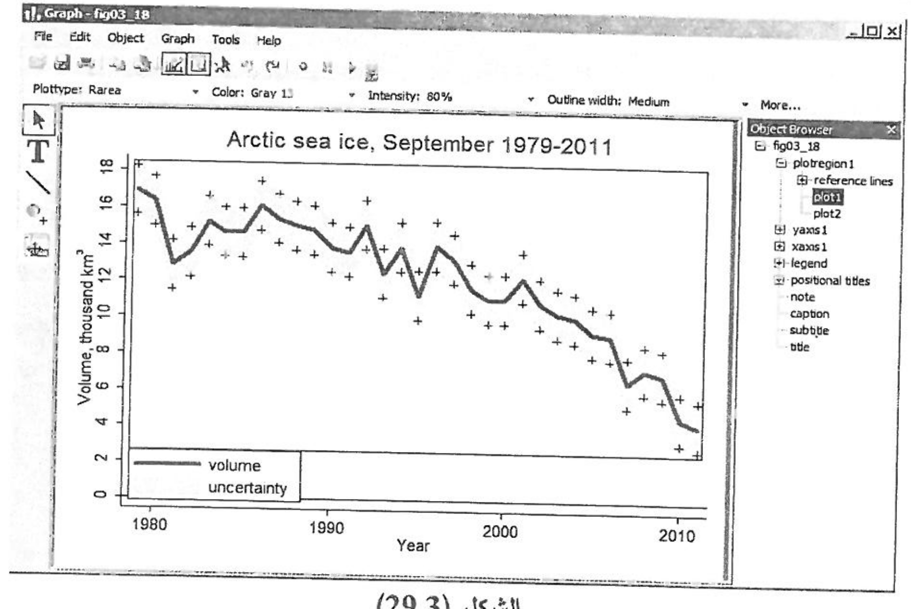
في نافذة الرسم البياني Graph Window اختر File > Start Graph Eidtor أو اضغط على أيقونة تحرير الرسم البياني سوف يتغير مظهر نافذة الرسم حيث سيظهر شريط أدوات في الهامش الأيسر للنافذة ومتصفح لعناصر الرسم في الجانب الأيمن، كما يظهر في الشكل (28.3) شريط أدوات يحتوي على أداة المؤشر لاختيار أجزاء من الرسم البياني وأدوات أخرى لإضافة نصوص أو خطوط، كما يمكن تحرير خطوط شبكة الرسم البياني، ويعرض متصفح عناصر الرسم قائمة لمحتويات الرسم البياني وتظهر علامة + بجانب بعض العناصر وبالنقر على علامة + سوف تتوسّع القائمة لتعرض عناصر أخرى داخلها، ويمكننا اختيار أي عنصر بالنقر عليه أو النقر على اسم العنصر في المتصفح في يمين النافذة (وهو الأسهل في الرسومات البيانية المعقدة).



الشكل (28.3)

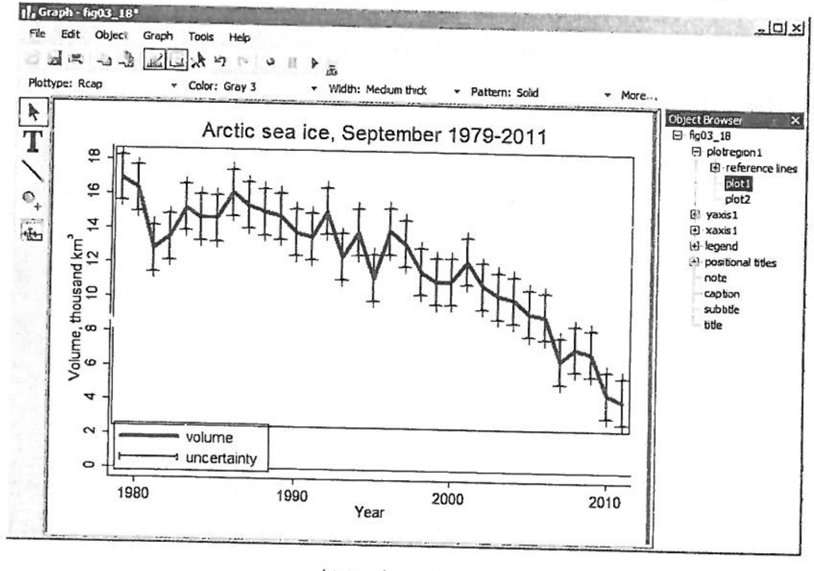
في الرسم البياني "Plot1" يشير إلى ناحية الرسم **twoway rarea** وهو تنسيق المنطقة الرمادية التي تمثل مدى الخطأ، والنقر على النطاق الرمادي (أو النقر على plot1 تحت القائمة plotregion1 بالمتصفح) لاختيارها سوف يتم تنشيط Plot1 في قائمة المتصفح، وفي الرسم البياني نفسه. إن عملية اختيار أي عنصر في الرسم سوف يفتح شريط أدوات علائقي فوق الرسم البياني مباشرة، وهذا الشريط يعطي معلومات عن خصائص العنصر الذي تم اختياره.

وفي هذا المثال يمكننا رؤية أن plot1 منطقة المدى الملونة باللون الرمادي Gray 13 وهو داكن بنسبة 80% وذو عرض متوسط، إذا قمنا بالنقر على ...More في شريط الأدوات العلائقي، يمكننا الحصول على خيارات أكثر للتحكم في خصائص العنصر الذي تم اختياره.



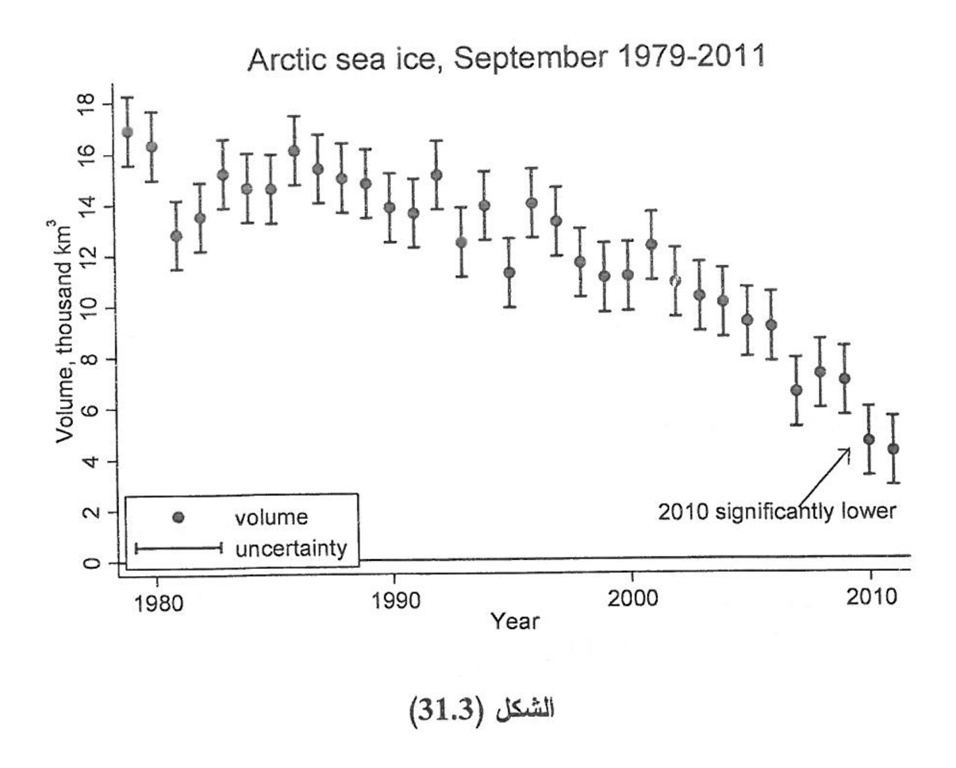
الشكل (29.3)

تغيير نوع الرسم البياني plottype من Rarea (منطقة المدى) إلى Rcap (منطقة مخططة) واللون من Gray 3 إلى Gray 13 (داكن أكثر) وعرض خط الرسم من Medium إلى Medium thick سوف يعطي الرسم شكلاً جديداً (الشكل 30.3).



الشكل (30.3)

في الرسم البياني Plot2 يشير إلى خط الرسم البياني **twoway line** والذي تم تركيبه على المنحنى الرئيس في الشكل (18.3)، إذا قمنا باختيار plot2 في متصفح عناصر الرسم البياني Graph Editor الموجود في الجانب الأيمن للنافذة، ثم قمنا بتغيير نوع الرسم plottype من Line إلى Scatter واللون إلى Gray 5 سوف نحصل على الشكل (29.3)، الرسم البياني النهائي يتضمن سهماً مع عبارة "2010 significantly lower" تم إضافتها بواسطة استخدام أداة T (نص) وأداة السهم، وهما من ضمن الأدوات الموجودة في شريط الأدوات بالجانب الأيسر للنافذة.



الشكل (31.3)

وعموماً، فإن نافذة محرر الرسم البياني Graph Editor لها خاصية تعديل الرسم البياني بعدة طرق يمكن التحكم بها باستخدام الأمر الأصلي **graph**، ولكن لا يمكننا عمل بعض الأشياء الأخرى مثل نقل نقطة بيانات معينة في الرسم بالرغم من أنه يمكننا إضافة أو حذف علامات جديدة في أي موضع

بالرسم. ومن ناحية أخرى، فإنه من السهل جداً تغيير خصائص العلامـات والخطوط وتوصيف المحاور والعناوين، كما يمكننا أيضاً إخفاء عناصر في الرسم البياني وجعلها غير مرئية، ويجب ملاحظة أن أي تغييرات في نافـذة محرر الرسم البياني Graph Editor تصبح دائمة عند حفظها.

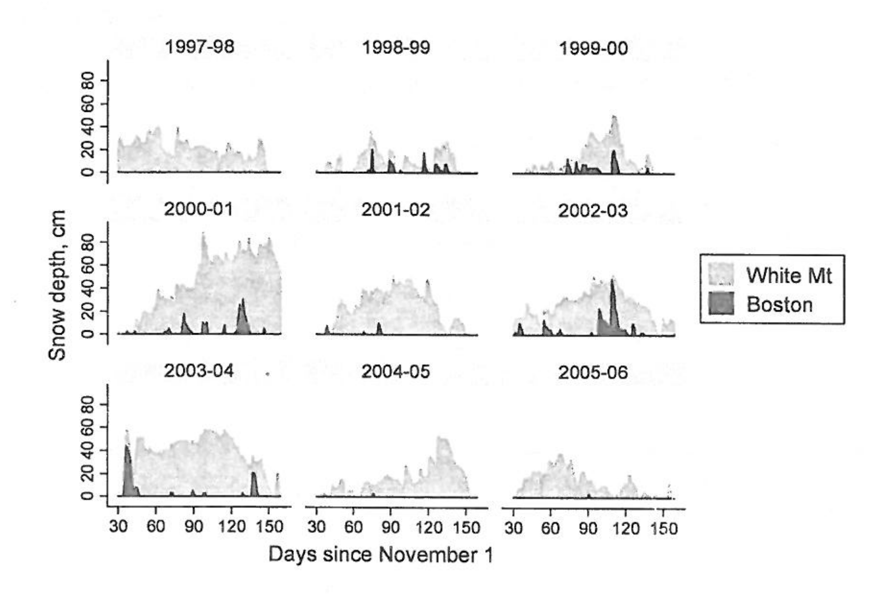
إبداعات في الرسم البياني : Creative Graphing

الكاتب Edward Tufte في كتبه المؤثرة والرائعة حول التمثيـل البيـاني للبيانات (1990، 1997، 2001، 2006) نصح بوضع مجهودات أكبر عند تصميم أي رسومات بيانية لتمثيل أي نوع من البيانات. يعتقـد الكاتـب أن عرض تصميمات رائعة وواضحة، يعطي للقارئ مساحة للمقارنة، وفحـص العلاقات بين مختلف المتغيرات، فمستخدمو برنامج ستاتا هم أناس عاديون، وستاتا يوفر مجموعة من الأدوات المرنة للتمثيل المرئي للبيانات المعقدة مما يسمح بتطوير الرسومات البيانية البسيطة أو إعادة تنظيمها بشكل جمالي أكثر لتظهر بأشكال جديدة.

أحد الموضوعات التي تناولها Edward Tufte هـو قيمـة المـضاعفات الصغيرة، وهي عبارة عن مجموعة من الرسومات البيانية الـصغيرة التـي تدمج في شكل واحد لتسهيل عملية المقارنة، فالأمر **graph** مع الخيـار **by()** يمكنه إنشاء مثل هذه الأشكال بطريقة رائعة، الشكل (32.3) يوضح بيانيـاً عمق الثلوج في فصل الشتاء خلال فترة زمنية معينة لمنطقتين الأولى هـي قرية في الجبال البيضاء بولاية نيو هامبشير بالولايات المتحدة والثانية مدينة بوسطون التي تبعد 225 كيلومترا جنوباً (البيانات بـالملف *whitemt1.dta*)، عمق الثلوج تم قياسه على أساس يومي في المنطقتين، وهذه البيانات تغطي تسعة فصول شتاء متتابعة في الفترة من شـتاء 1997-1998 إلـى شـتاء 2005-2006 والمتغير *dayseason* يقوم بعد الأيام من 1 نوفمبر لكل فصل شتاء، والمتغيران *mtdepth* و*bosdepth* يقيسان عمق الثلوج بالسنتيمتر في كل من الجبال البيضاء ومدينة بوسطون على التوالي، المتغيـر *season* يحـدد فصول الشتاء للفترة من 1997-1998 إلى الفتـرة 2005-2006، باتبـاع الأمر **twoway area** نقوم بإنشاء رسم بياني للمتغيـرين *mtdepth* و*bosdepth*

مع المتغير *dayseason* ونقــوم بتحــديد الألوان لتكــون رماداً فاتحاً ورمادياً داكناً (gs11, gs5) والتصميم يكون 3×3 لكل فصل شتاء ونحدد بــأن مربــع شرح الرسم البياني يكون في عمود واحد يتم وضعه في موضع الــساعة 3، والخيار **symxsize(*.3)** يوفر مساحة في الرسم البياني وذلك بوضع الرمــوز في مربع شرح الرسم البياني ليكون عرضــها 30% بــدلاً مــن العــرض الافتراضي.

```
.graph twoway area mtdepth bosdepth dayseason
 if dayseason>29 &dayseason<160,
 bcolor(gs11 gs5) ytitle("Snow depth, cm")
 by(season, rows(3) note("")
 legend(position(3)))
 xlabel(30(30)150) ylabel(0(20)80)
 legend(cols(1) label(1 "White Mt") label(2
 "Boston")
 symxsize(*.3))
```



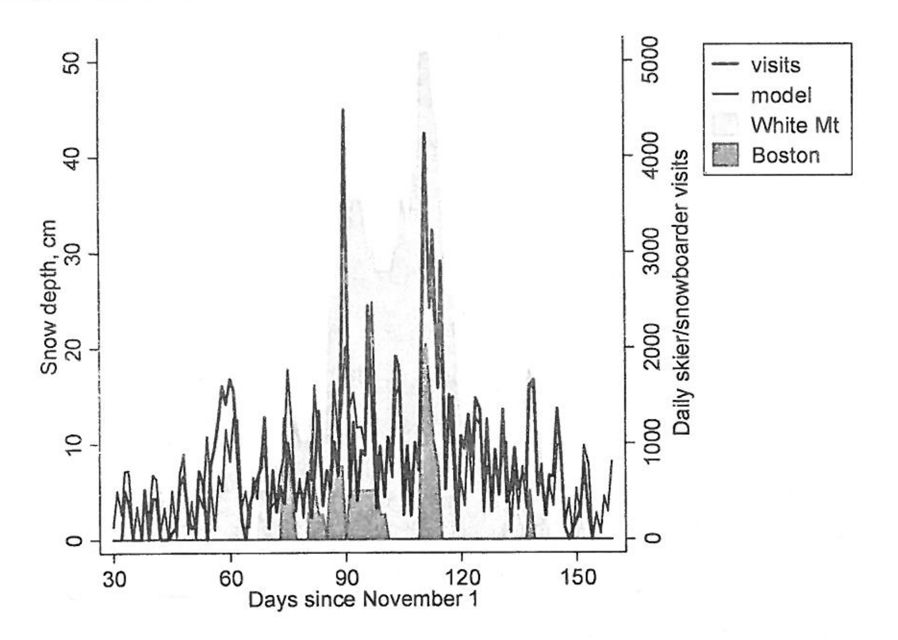
الشكل (32.3)

الشكل (32.3) يعرض الأجواء اليومية خلال تسعة فصول شتاء بولاية نيو إنجلاند، موضحاً كيف أن عمق الثلوج يتغير في مكانين مختلفين وفترات

زمنية مختلفة، ففي الفترة من 2000–2001 إلى 2003–2004 يبدو أن هطول الثلوج كان كثيفاً في الجبال مع بعض العواصف القوية في مدينة بوسطون، أما خلال الفترة 1998–1999 كانت الثلوج أعلى في الجبال مع وجود فترات بدون أي وجود للثلوج على الأرض.

البيانات التي تم تمثيلها في الشكل (32.3) تم جمعها لدراسة كيفية تأثير الظروف الجوية بولاية نيوإنجلاند على مستوى الحضور في مناطق التزلج (دراسة Hamilton et al. 7002, 2003)، وبما أن مناخ الشتاء بولاية نيوإنجلاند أصبح أكثر دفئاً في العقود الأخيرة، فإن انخفاض مستوى الثلوج أصبح هو السمة الغالبة، وارتفاع الحرارة يعتبر مشكلة بيئية لها أبعادها التي تؤثر على نواح أخرى مثل التزلج خلال فصل الشتاء، مناطق التزلج تتأثر ليس فقط بظروف تساقط الثلوج في المناطق المحلية وإنما تتأثر أيضاً بمستوى الثلوج في المدن البعيدة مثل مدينة بوسطون حيث يعيش العديد من المتزلجين. الشكل البياني التالي (33.3) يركز على فصل شتاء واحد لسنة 1999 – 2000 (ملف البيانات *whitemt2.dta*)، حيث يبدأ بنفس عمق الثلوج في حواف الجبال والتي ظهرت في أعلى اليمين بالشكل (32.3).

الشكل (33.3) يضع بيانات الجليد في حواف الجبال (منطقة الرسم البياني **twoway area**) فوق رسم بياني خطي **line** موضحاً عدد المتزلجين وعدد زيارات المتزلجين في كل يوم في منطقة تزلج واحدة في الجبال البيضاء بالقرب من مكان جمع بيانات عمق الثلوج، تم إنشاء رسم بياني يوضح عدد الزيارات (*visits*) وعدد الزيارات المتوقع الذي تم حسابه بنموذج سلاسل زمنية (*model*)، النموذج تم شرحه بدراسة Hamilton et al. (2007) يتوقع الحضور اليومي كدالة للعوامل الدورية الأسبوعية مع الجو وظروف الثلوج في الجبال ومدينة بوسطون، الأمر **graph** يقوم بإنشاء الشكل (33.3) مخصصاً المحور *y* بالجانب الأيسر لتمثيل بيانات عمق الثلوج بالسنتيمتر (*mtdepth, dosdepth*) والمحور *y* في الجانب الأيمن لتمثيل عدد الزوار، والنموذج الخاص بالتنبؤ (*visits, model*).



الشكل (33.3)

يجب ملاحظة أن الإعدادات الدقيقة للخيارات **yscale(range())** و **ylabel()** لكل من الرسمين البيانيين في الشكل (33.3) حيث استطعنا أن نطابق بين المقاييس، بحيث إن خطوط الشبكة الأفقية في الرسم تمثل مقاييس متغيرين اثنين في نفس الوقت وهذا ليس عملياً لكل أنواع البيانات، ولكن يمكنه تطوير الرسم بزيادة وضوحه وإظهاره لمقاييس مختلفة على المحاور العمودية على يمين ويسار الرسم البياني.

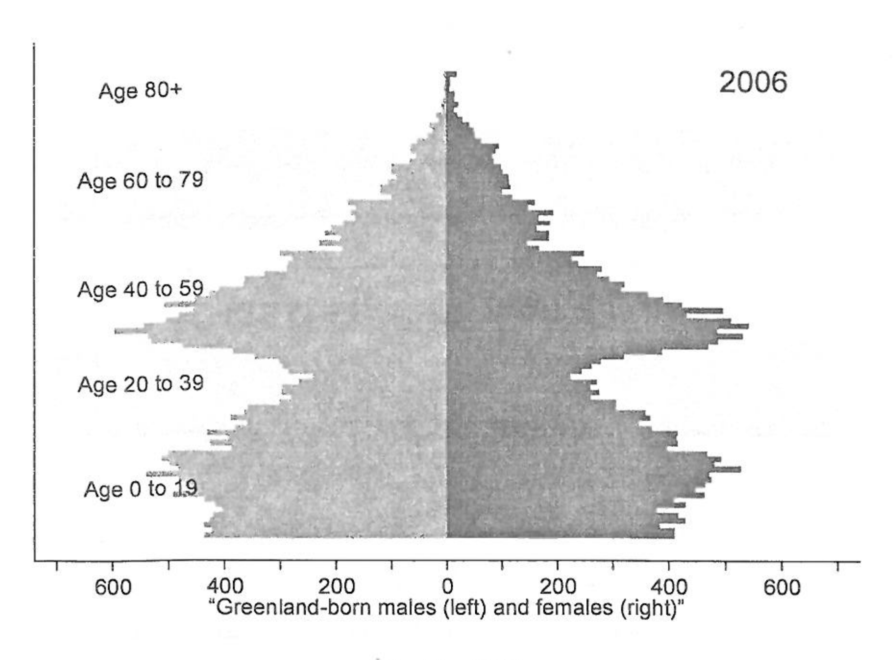
```
.graph twoway area mtdepth bosdepth dayseason,
yaxis(1)
ytitle("Snow depth, cm", axis(1)) bcolor(gs12
gs6)
ylabel(0(10)60, axis(1))
|| line model visits dayseason, yaxis(2)
lwidth(medthin medthick)
ylabel(0(1)3, axis(2)) lcolor(gs1 gs0)
|| if dayseason>29 &dayseason<160,
r2("Daily skier/snowboarder visits")
xlabel(30(30)150)
xtitle("Days since November 1")
legend(rows(4) position(2) order(4 3 1 2)
label(1 "White Mt")
```

```
label(2 "Boston") label(3 "model") label(4
"attend")
symxsize(*.3))
yscale(range(0,51) axis(1)) ylabel(0(10)50,
axis(1) grid)
yscale(range(0,5100) axis(2))
ylabel(0(1000)5000, axis(2))
```

أعلى قمتين في الرسم البياني في زيارات منطقة التزلج كانتا خلال فترة عطلة المدارس والتي تأتي بالتصادف مع هطول الثلوج في مدينة بوسطون. الدراسة الأصلية قامت باختبار وتأكيد التأثير المعنوي لهذه العوامل، وسوف يكون من الأسهل تمثيل البيانات بعدة أشكال بيانية ودمجها في شكل واحد، كما تم في الشكل (32.3) بدلاً من استخدام شكل واحد كما حدث في الشكل (33.3).

أهرامات عدد السكان والتي تُستخدم بشكل كبير بواسطة علماء الجغرافيا السكانية لتمثيل تركيبة السن والجنس للسكان، وهذه الأهرامات ليست ضمن أنواع الرسم البياني ببرنامج ستاتا، ولكن يمكن تركيبها بواسطة أعمدة بيانية أفقية من خلال استخدام الأمر **graph hbar**، وهناك العديد من الطرق الأخرى للقيام بذلك، الشكل (34.3) يوضح إحدى طرق إنشاء هذا النوع من الرسومات البيانية باستخدام هرم لمعدل المواليد بمقاطعة جرين لاند في الدينمارك والتي تقطنها أغلبية عرقية من الإسكيمو في سنة 2006 (المرجع: Hamilton and Rasmussen 2010) عدد الإناث في كل فئة عمرية تم الإشار إليه بواسطة عمود إلى يمين مركز الرسم، وعدد الذكور الذي في نفس الفئة العمرية بواسطة عمود إلى اليسار، المجموعات ذات الفئة العمرية 90 سنة كانت كبيرة لتوصيفها بشكل منفرد، لذا فقد تم تلوينها بلون بشريط رمادي لكل 20 سنة (0–19 سنة، 20–39 سنة وهكذا). على سبيل المثال، الرسم البياني يُشير إلى أنه في سنة 2006 معدل المواليد يتضمن 600 من الذكور أعمارهم 40 سنة، ولكن أقل من 500 من الإناث أعمارهن 40 سنة لنفس الفترة، وهذا يعكس اختلاف الجنس في صافي الهجرة الخارجية، البروز الرئيس في هذا الهرم يُظهر الارتفاع الكبير في معدل البالغين في الفئة العمرية من 35–49 سنة (الذين ولدوا في الخمسينيات والستينيات) متبوعاً بمجموعات من الشباب البالغين، كما يمكننا مشاهدة زيادة واضحة في عدد

المواليد الذين ولدوا في الثمانينيات والتسعينيات إلى البالغين من الزيادة الأولى في معدل المواليد، الفئة العمرية من 10-14 سنة تشمل مجموعة كبيرة من الأطفال.



الشكل (34.3)

هناك العديد من الطرق التي يمكن استخدامها لإنشاء الشكل (34.3) فصف البيانات (*greenpop.dta*) يحتوي على عدد من الذكور *males* والإناث *females* في كل الأعمار *age*، ولإنشاء رسم بياني يمثل الذكور على اليسار، نحن نحتاج إلى إنشاء متغير جديد يساوي عدد الذكور بإشارة سالبة.

```
.gen negmales=-males
```

يمكن إنشاء هرم بسيط بدون أي توصيفات وذلك بطباعة الأمر:

```
.graph hbar (sum) negmales females if year==2006,
over(age, descending gap(0) label(nolabel))
```

ولإظهار نطاقات رمادية في الخلفية بحيث تبرز نطاقات لكل 20 سنة، فإننا نحتاج إلى تحديد متغيرات وهمية هي *maleGRAY* و*femGRAY* حتى نملأ الرسم البياني بموجب أو ناقص 700 كما يلي:

```
.gen maleGRAY = -(700-males) if (age>=20
 &age<40)
|(age>=60 & age<80)
.gen femGRAY = 700-females if (age>=20 &age<40)
|(age>=60 &age<80)
```

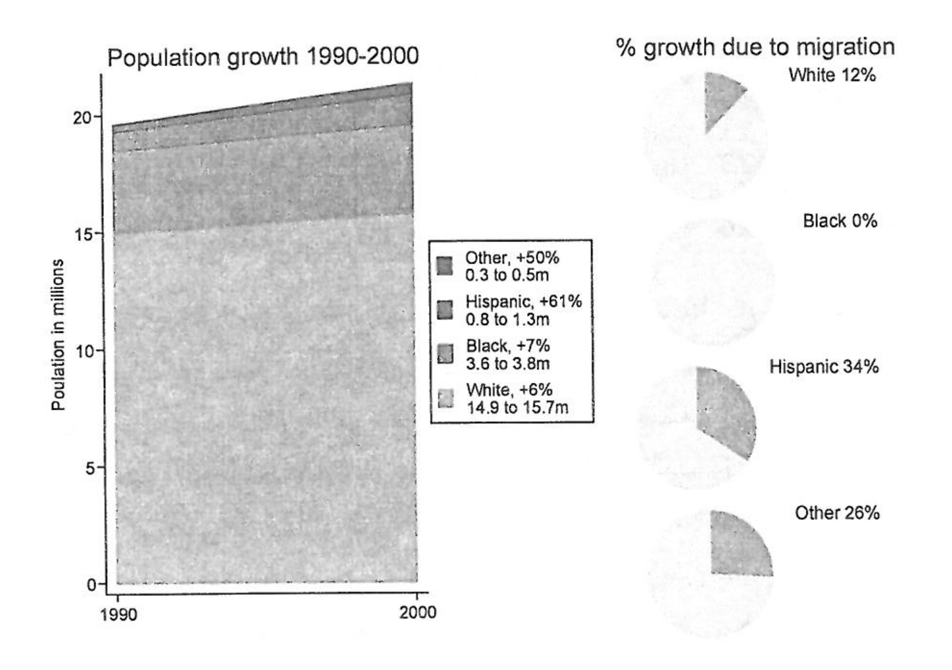
الشكل (34.3) يمكن رسمه الآن، وذلك بوضع المتغيرات *negmales, females, maleGRAY, femGRAY* في أعمدة أفقية مع نصوص لتوصيف المناطق الرمادية، كما يمكننا تطبيق توصيفات مثل "600" على 600- على المحور العمودي *y* حتى لا يظهر عدد الذكور سالباً في الرسم.

```
.graph hbar (sum) negmales females malGRAY
 femGRAY if
year==2006,
over(age, descending gap(0) label(nolabel))
ylabel(-600 "600" -400 "400" -200 "200" 0 200
 400 600)
ytick(-700(100)700, grid) legend(off) stack
ytitle("Greenland-born males (left) and females
 (right)")
bar(1, color(emidblue)) bar(2, color(maroon))
 bar(3,
color(gs14))
bar(4, color(gs14)) text(550 97 "2006",
 size(large))
text(-550 11 "Age 0 to 19")
text(-550 33 "Age 20 to 39") text(-550 53 "Age
 40 to 59")
text(-550 76 "Age 60 to 79") text(-550 95 "Age
 80+")
```

الشكل (35.3) يدمج خمسة رسومات بيانية مبسّطة مع نص ليعرض شكلاً بيانياً له بعض خصائص الجدول والشرح معاً. الشكل الجديد يعرض التغير في عدد السكان للفترة من 1990–2000 لمختلف المجموعات العرقية التي تعيش في المحافظات البعيدة بجنوب الولايات المتحدة (حسب بيانات التعداد السكاني للولايات المتحدة لسنة 2005) الجانب الأيسر للشكل (35.3) عبارة عن رسم بياني من نوع **twoway** وللقيام بعرض تغيرات عدد السكان تم تمثيل المتغيرات بيانياً لكل مجموعة عرقية *popwbho, popwbh*..الخ وهي تمثل مجاميع تم حسابها لكل مجموعة عرقية كما يظهر في الشكل البياني أدناه (البيانات موجودة بالملف *southmig1.dta*) معلومة إضافية مهمة لا تظهر

في منطقة الرسم البياني نفسه ولكن يمكن ملاحظتها من خطين تم توصيفهما لكل مجموعة عرقية في مربع شرح الرسم، فمثلاً يستطيع القارئ أن يـرى من خلال مربع شرح الرسم، أن عدد السكان الذين تمتد أصولهم من أمريكا اللاتينية Hispanic قد زاد في الولايات الجنوبية بنسبة 61% خلال هذا العقد من حوالي 800,000 نسمة إلى 1.3 مليون نسمة وأصبحت نـسبتهم أكثـر وضوحاً مقارنة مع باقي المجموعات العرقية الأخرى.

```
.graph twoway area popwbho popwbh popwb popw
year,
legend(rows(4) position(3) symxsize(3)
label(1 "Other, +50%" "0.3 to 0.5m")
label(2 "Hispanic, +61%" "0.8 to 1.3m")
label(3 "Black, +7%" "3.6 to 3.8m")
label(4 "White, +6%" "14.9 to 15.7m"))
xlabel(1990 2000) xtitle("")
ylabel(0(5)20,angle(horizontal) grid)
ytitle("Population in millions")
title("Population growth 1990-2000")
```



الشكل (35.3)

الجانب الأيمن للشكل (35.3) يتكون من أربعة أشكال دائرية تعرض النسب المئوية للنمو السكاني نتيجة الهجرة، كل دائرة تم رسمها بشكل منفصل باستخدام البيانات الموجودة بالملف *southmig2.dta* فمثلاً آخر شكل دائري يوضح أن 12% من العرق الأبيض كان نموه نتيجة الهجرة، المتغيرات التي تظهر في الرسم البياني تمثل صافي الهجرة (*netmig_w* عبارة عن المجموع الكلي للهجرة الداخلية ناقص الهجرة الخارجية) ونمو عدد السكان نتيجة الزيادة الطبيعية (*nonmig_w* عدد المواليد ناقص عدد الوفيات).

```
.graph pie nonmig_w netmig_w,
legend(off) pie(1, color(dkorange)) pie(2,
color(gs13))
title("White 12% ", position(2))
```

كل رسم بياني دائري يتم حفظه باسم، مثل *pie_white.gph* وبعد إنشاء وحفظ الأشكال الأربعة الدائرية يمكن دمجها معاً باستخدام الأمر **.graph combine**

```
.graph combine pie_other.gph pie_hisp.gph
pie_black.gph pie_white.gph,
imargin(tiny) rows(4)
title("% growth due to migration") fxsize(40)
```

الخيار fxsize(4): يقوم بدمج الأشكال الدائرية الأربعة في شكل واحد ويستخدم 40% فقط من العرض المتوافر، وبالتالي عند دمجه مع الجانب الأيسر للشكل (35.3) فإن الرسم البياني الدائري سوف يأخذ أقل من نصف العرض الكلي للشكل.

هذه الأمثلة توضح بعض الإمكانات لتصميم الرسومات البيانية ببرنامج ستاتا بدمج العناصر مع بعضها بطريقة أكثر وضوحاً.



الفصل الرابع

بيانات الدراسات الاستقصائية

Survey Data

هذا الفصل يعتبر مقدمة مختصرة للعمل مع بيانات الدراسات الاستقصائية باستخدام برنامج ستاتا، القرّاء غير المهتمين بالدراسات الاستقصائية، يمكنهم الانتقال إلى الفصول الأخرى، فلن تؤثر عليهم عدم قراءتهم لهذا الفصل. ومن ناحية أخرى، المهتمون بالعلوم الاجتماعية والتي يعتبر فيها استخدام البيانات الاستقصائية أمرًا في غاية الأهمية، قد يكون من المفيد لهم إلقاء نظرة سريعة عن كيفية التعامل مع مثل هذه البيانات. أما الطرق المتقدمة في تحليل البيانات، فسوف تأتي لاحقاً في الفصول القادمة.

بحوث الدراسات الاستقصائية تركز بشكل كبير على رسم استدلالات صحيحة حول السكان، ونظرياً العينة العشوائية البسيطة هي أفضل وسيلة لمثل هذا النوع من الدراسات، ولكن في العادة، فإن هذه الوسيلة مكلفة جداً، وبدلاً من استخدام العينة العشوائية البسيطة يمكن استخدام طرق أخرى أسهل، ولكنها أحياناً تستخدم استراتيجيات معينة معقدة لتحقيق أهداف معينة، وللحصول على تمثيل معقول لمجتمع الدراسة. كما أنها قد تتطلب إجراء بعض التعديلات بعد إجراء عملية المعاينة. وحيث إن الطرق الإحصائية المعيارية تقوم على فرضية أن العينة عشوائية بسيطة لذلك فنحن نحتاج إلى طرق متخصصة لتصميم بيانات الدراسات الاستقصائية التي يمكن أن تأخذ في الاعتبار الحصول على معلومات عن إجراءات المعاينة.

برنامج ستاتا من البرامج القوية في تحليل بيانات الدراسات الاستقصائية، معتمداً على مدخل موحّد يشمل مجموعة كبيرة من طرق التحليل الإحصائي، كل هذه الطرق تعمل من خلال تعريف أساسي لتركيبة

العينة والتي يمكن أن تتضمن أوزاناً احتمالية لتقليل التحيز عند اختيار العينة. وتصميم العينة يمكن أن يتضمن أيضاً تعقيدات أخرى مثل تحديد الطبقات العنقودية متعددة المراحل، والمجتمعات المحددة، في مستوى واحد أو أكثر، والمعاينة الإرجاعية، والأوزان التكرارية، ومابعد تحديد الطبقات؛ وعند تحديد عناصر التصميم الرئيسة في الدراسة الإستقصائية (باستخدام الأمر **svyset**)، سوف يتم حفظ مجموعة من البيانات باستخدام المعلومات المعطاة، وبعد ذلك يتم التحليل باستخدام الأمر **svy** الذي سيقوم بتطبيق الأوزان ومعلومات التصميم الأخرى تلقائياً كما هو مطلوب.

أغلب إجراءات الدراسات الاستقصائية يمكن إجراؤها باستخدام القوائم والعديد من القوائم الفرعية وذلك باختيار Statistics > Survey data analysis. وللحصول على معلومات أكثر حول أوامر الدراسات الاستقصائية، قم بطباعة الأمر **help survey** كما أن دليل المستخدم *Survey Data Reference Manual* يشرح العديد من الأمثلة والتفاصيل التقنية لجميع أوامر ستاتا المتعلقة بالدراسات الاستقصائية. المراجع الأخرى المفيدة تتضمن كتاباً عن المعاينة للكاتبين Levy and Lemeshow (1999) وكتاب عن تحليل بيانات الدراسات الاستقصائية مع أمثلة عن الإحصائيات الحيوية للكاتبين Korn and Graubard (1999)، كما أن Sul and Forthofer (2006) يقدمان نظرة مختصرة عن المسائل الرئيسة في تحليل بيانات الدراسات الاستقصائية. وللاطلاع عن نقاش مفصّل عن مشاكل الأسئلة المفتوحة بالدراسات الاستقصائية انظر كتاب Moore (2008).

أمثلة عن الأوامر : Example Commands

```
.svyset _n [pweight = censuswt]
```

يقوم بتحديد البيانات على أنها بيانات دراسة استقصائية مع أوزان احتمالية (نسبي إلى الاحتمال العكسي للاختيار) تم إعطاؤها بواسطة المتغير *censuswt*، المحدد _n الذي يحدد المشاهدات الفردية (افتراضي) كوحدات معاينة رئيسة.

```
.svyset _n [pweight = censuswt], strata(district)
```

يُحدد بأن البيانات هي بيانات دراسة استقصائية مأخوذة من عينة طبقية ذات مرحلة واحدة، حيث تم تقسيم المجتمع إلى طبقات وأفراد، وتم أخذ عينة مستقلة من كل طبقة، وفي هذا المثال فإن المتغير *district* يحدد الطبقات، والمتغير *censuswt* يحدد الأوزان الاحتمالية.

```
.svyset school [pweight = finalwt], fpc(nschools)
|| _n, fpc(nstudents)
```

يحدد أن البيانات هي بيانات دراسة استقصائية من عينة عنقودية ذات مرحلتين، في المرحلة الأولى تم اختيار المدارس بطريقة عشوائية، لذا فإن المدارس هي وحدة المعاينة الأولية. أما في المرحلة الثانية، فتم خلالها اختيار الطلبة بطريقة عشوائية من المدارس التي تم اختيارها في المرحلة الأولى، وفي كل مرحلة تم تحديد تصحيحات المجتمع المحدد (FPSs)، فالمتغير *nschools* يمثل مجموع عدد المدارس في المجتمع، والمتغير *nstudents* يمثل مجموع عدد الطلبة في كل مدرسة.

```
.svy: tabulate vote, percent miss ci
```

الحصول على جدول بالنسب المرجحة وحدود الثقة للمتغير *vote* يتضمن القيم المفقودة بناء على بيانات دراسة استقصائية **svyset**.

```
.svy: tabulate vote gender,column pearson lr wald
```

الحصول على جدول تقاطعي لأوزان المتغير *vote* مع المتغير *gender* مع نسبة كل عمود للمتغير *gender*، كما أن الجدول سوف يعرض نتائج اختبار مربع كاي تربيع لبيرسو، ومعدل الأرجحية كاي تربيع وإحصائيات اختبار وولد Wald test.

```
.svy: regress y x1 x2 x3
```

يعرض انحدار المتغير *y* بالدراسة الاستقصائية مع ثلاثة متغيرات تنبؤية وهي *x1, x2, x3*. وللحصول على قائمة كاملة لأنواع أوامر الانحدار وطرق التقدير في الدراسات الاستقصائية قم بطباعة الأمر **help svy estimation**.

```
.svy, subpop(voted): regress y x1 x2 x3
```

يقوم بتحليل الانحدار للدراسة الاستقصائية باستخدام مجتمع ثانوي تم تعريفه بواسطة 1 والقيم {0,1} للمتغير *voted*، اختيار مجموعة ثانوية من البيانات يتم بطريقة اعتيادية من خلال استخدام محددات مثل **if** أو **in** وهذا لا يعتبر مناسباً عند تحليل الدراسات الاستقصائية، ولكن بدلاً من ذلك يتم استخدام الخيارات **svy** و**subpop()**.

```
.svy: mean age, over(gender)
```

يقوم هذا الأمر بحساب المتوسط المرجح، وفترات الثقة للمتغير *age* لفئات المتغير *gender*.

تحديد بيانات الدراسة الاستقصائية : Declare Survey Data

منذ سنة 2001 قام مركز ولاية جرانيت لاستطلاعات الرأي بجامعة نيوهامبشير، بإجراء دراسات استقصائية عبر الهاتف عدة مرات في كل سنة، وكل دراسة قامت بالاتصال بعينة جديدة تتكون من 50 شخصاً وسؤالهم عن آرائهم وعن خلفياتهم. نتائج استطلاعات الرأي السياسية تجذب اهتمام الرأي العام كل أربع سنوات خاصة خلال فترة الحملات الانتخابية الرئاسيّة بولاية هامبشير، الملف *Granite2011_6.dta* يحتوي على أسئلة لدراسة استطلاعية قام بها مركز جرانيت لعدد 516 شخصاً في يونيو 2011.

```
.use C:\data\Granite2011_6.dta, clear
.describe, short
```

```
Contains data from C:\data\Granite2011_6.dta
  obs:           516                         New Hampshire, Granite State
                                               Poll -- June 2011
 vars:            33                         2 Jul 2012 06:11
 size:        19,608
Sorted by:  respnum_
```

كما يحدث في أي دراسة استقصائية فإن تصميم المعاينة، ونمط الردود ربما يؤديان إلى الحصول على بيانات تختلف عن المجتمع المستهدف. فمثلاً بيانات التعداد السكاني توضح بأن أقل من 52% من البالغين في الولاية هم من الإناث، ولكن النساء يمثلن نحو 55% من نسبة العينة.

```
.tab sex
```

Gender	Freq.	Percent	Cum.
Male	234	45.35	45.35
Female	282	54.65	100.00
Total	516	100.00	

لتعويض التحيز البسيط في العينة، تقوم بحوث الدراسات الاستقصائية بحساب الأوزان الاحتمالية. فبعض هذه الأوزان يتم حسابها من خلال المقارنة بين خصائص العينة والمجتمع مثل الجنس في العينة أعلاه، والحسابات الأخرى يتم القيام بها بناءً على خصائص تصميم العينة، بالنسبة للباحثين بمركز جرانيت فإنهم يقومون بتعريف المتغير *censuswt* على أنه عبارة عن تجميع لحجم أرباب الأسر، وأرقام الهواتف، والجنس، والديانة بالولاية. وقيم المتغير *censuswt* في استطلاع الرأي لشهر يونيو 2011 كان متوسطه 1، ولكن المتوسط يتغير من 0.16 (هناك بعض المشاهدات تم إعطاؤها أوزاناً مرجحة منخفضة لتعويض التمثيل المرتفع في العينة) إلى 2.19 (تم إعطاء أوزان مرجحة مرتفعة لتعويض التمثيل المنخفض في العينة).

```
.summarize censuswt
```

Variable	Obs	Mean	Std. Dev.	Min	Max
censuswt	516	.9991743	.4601123	.1603937	2.194549

الأمر svyset: يحدد أن البيانات الموجودة هي بيانات دراسة استقصائية، مع إعطاء أوزان احتمالية بواسطة المتغير *censuswt*، وحفظ هذه البيانات، ثم حفظ هذه المعلومات يعتبر مهماً بالرغم من إمكانية استخدام الأوزان في أي تحليل إحصائي عند الحاجة إلى ذلك في تحليل معين، ما عدا ذلك فإن هذه البيانات لن تتغير.

```
.svyset _n [pweight = censuswt]

      pweight: censuswt
          VCE: linearized
  Single unit: missing
     Strata 1: <one>
         SU 1: <observations>
        FPC 1: <zero>
```

```
.save, replace
.svydescribe
    Survey: Describing stage 1 sampling units

         pweight: censuswt
             VCE: linearized
     Single unit: missing
        Strata 1: <one>
            SU 1: <observations>
           FPC 1: <zero>

                                          #Obs per Unit
                                   ----------------------------
 Stratum   #Units     #Obs        min       mean        max
 -------- -------- --------    -------- ---------- ----------
       1      516      516          1        1.0          1
 -------- -------- --------    -------- ---------- ----------
       1      516      516          1        1.0          1
```

عند تحديد بيانات الدراسة الاستقصائية باستخدام الأمـــر svyset، فـــإن الأوامر التي تبدأ بـــ svy سوف تقوم بإجراء الحسابات الإحصائية مستخدمة معلومات وزن الدراسة الاستقصائية، وبعد ترجيح تمثيل الجنس بالعينة ليكون أقرب للقيمة المتوقعة في المجتمع.

```
.svy: tab sex
(running tabulate on estimation sample)

Number of strata   =         1        Number of obs      =       516
Number of PSUs     =       516        Population size    = 515.57392
                                      Design df          =       515

  ----------------------
    Gender | proportions
  ----------+-----------
      Male |       .4965
    Female |       .5035
           |
     Total |           1
  ----------------------
    Key:  proportions  =  cell proportions
```

العديد من أوامر ستاتا من الجداول البسيطة إلى النمـــاذج الإحـــصائية، تسمح بإضافة : svy قبل الأمر. فمثلاً يمكننا حـــساب الانحـــدار المنطقـــي المرجح (في الفصل 9) لوجهات النظر الفردية حـــول التغيـــر المنـــاخي –

المتعلقة بالمتغير *warmop2* – مع مستوى تعليم المشارك في الدراسة وحزبه السياسي من خلال الأمر التالي:

```
.svy: logit warmop2 educ party
```

للحصول على قائمة بالاحتمالات التحليلية، قم بطباعة الأمر **help survey**، الأمر **svyset** يمكنه أن يحدد معلومات أكثر على أنها معلومات دراسة استقصائية بطرق أخرى مختلفة عن تلك التي سبق أن رأيناها في الأمثلة السابقة، فخيارات الأمر **svyset** تسمح لنا بإنشاء تصميمات معقدة، تتضمن معاينة عنقودية متعددة المراحل، وطبقية وتصحيح المجتمع المحدد وطرق بديلة لتقدير التباين وإضافة طبقات جديدة. وللحصول على قائمة كاملة بالخيارات المتوافرة وتركيباتها، قم بطباعة الأمر **help svyset**. كما أن دليل المستخدم *Survey Data Reference Manual* يشرح أمثلة وتقنيات عن هذا الموضوع بشكل أكثر تفصيلاً.

تصميم الأوزان : Design Weights

الجزء السابق ركّز على تعريفات الأوزان كمسلمات. والعديد من مستخدمي البيانات يبدأون عملهم ببيانات استقصائية كاملة تم حساب أوزانها مسبقاً. هذا الجزء والأجزاء القادمة، سوف تعرض أمثلة توضح كيفية حساب الأوزان.

الباحثون الذين يستخدمون بيانات الدراسات الاستقصائية يطبقون الأوزان الاحتمالية لضبط التحيز في طرق المعاينة، فقد يظهر التحيز نتيجة عاملين اثنين هما خصائص مُتعمّدة في تصميم المعاينة أو خصائص مضافة بشكل غير مقصود أثناء عملية جمع البيانات، وكلا العاملين يؤديان إلى الحصول على عينة غير ممثلة للمجتمع، ولاتعطي صورة واقعية عن تقلبات العينة وخصائص المجتمع.

بالنسبة لمركز جرانيت لاستطلاع الرأي، فإن الباحثين يقومون بالاتصال بعينة عشوائية من سكان نيوهامبشير عن طريق الهاتف، ونظرياً أرقام الهاتف العشوائية يمكن أن تنتج عينة عشوائية من سكان الولاية، فعند إجراء

استطلاعات عن الانتخابات أو أي موضوع آخر، فإن الباحث يريد تعميم نتائجه ليس فقط على السكان الذين تم الاتصال بهم ولكن على كل الناخبين الذين يعيشون في الولاية. فبعض السكان هم من البالغين فقط، وبعضهم الآخر مختلط. بين المجيبين عن الاتصالات الهاتفية لاستطلاع يونيو 2011 هناك نحو 29% قالوا إنهم يعيشون في بيت به شخص بالغ واحد فقط. الإجابات في هذا المثال محددة: "واحد، اثنين، ثلاثة أو أكثر". وهذا شيء عملي للمقارنة بين الأوزان الترجيحية، هناك 503 أشخاص فقط من أصل 516 قاموا بالإجابة عن سؤال عدد البالغين في البيت، سوف نعود لاحقاً للذين لم يجيبوا عن هذا السؤال وعددهم 13 شخصاً.

```
.tab adults
```

# adults in household	Freq.	Percent	Cum.
1	148	29.42	29.42
2	273	54.27	83.70
3+	82	16.30	100.00
Total	503	100.00	

بالرغم من أن 29% من أفراد العينة يعيشون في بيت به شخص بالغ واحد، فإنه من الخطأ توقّع نفس النسبة لكل سكان ولاية هامبشير. ولاختيار شخص واحد بطريقة عشوائية عند إجراء الاتصالات الهاتفية، الباحث الذي قام بالاتصال سوف يسأل عن شخص بالغ في البيت للحديث معه، أو سوف يقوم بالاتصال في وقت آخر عند عدم الرد على الهاتف، هذا يؤدي إلى أن البيوت التي يعيش فيها شخص بالغ واحد أقل احتمالاً بثلاث مرات أن يدخلوا في العينة مقارنة بالبيوت التي يعيش فيها ثلاثة بالغين أو أكثر. الجدول أعلاه، يوضح بأنه يجب عد المكالمات الهاتفية لكل البيوت التي يجب أن تكون على الأقل (1×148)+(2×273)+(3×82)=940 شخصاً بالغاً. في هذه العينة الوهمية الذين يعيشون في بيت به شخص بالغ واحد يمثلون 148÷940 أو نسبة 16% وهي أقل بكثير من نسبة 29% بالجدول أعلاه.

أوزان الدراسات الاستقصائية تعتبر طريقة لتصحيح التحيز في العينـــة، كما أنها تساعد في الحصول على نتائج أكثر واقعية. في هذا المثال الأوزان مهمة ليس فقط لشرح عدد السكان في الولاية، وإنما أيضاً مهمة لأغــراض أخرى مثل السلوك الانتخابي الذي قد يرتبط بحجم السكان. فوجود شــخص بالغ واحد في المنزل، ربما يرتبط بنسبة كبيرة من كبار السن الذين يعيشون وحدهم، ووجود شخصين بالغين في المنزل يعني وجود العديد مــن الأســر الشابة، ووجود عدة أشخاص بالغين دلالة على أن العائلات مسنّة أو وجــود بالغين شباب مع أصدقائهم يعيشون في المنزل.

الأوزان الاحتمالية تتناسب مع معكوس احتمال الاختيار. ففــي المثــال أعلاه الاحتمال الشرطي لاختيار شخص معين من منزل به شــخص بــالغ واحد (بافتراض أننا قمنا بالاتصال بذلك المنزل) يساوي واحــد، واحتمــال اختيار شخص معين من منزل به شخصان بالغان يساوي 2/1 ومن منــزل به ثلاثة أشخاص بالغين 3/1. وإذا قمنا باستخدام معكوس هذه الاحتمــالات 1، 2، 3 كأوزان، فإن العينة سوف تتضمن 940 شخصاً وهمياً ولكن هــذا سوف يقود إلى مجاميع غير صحيحة، ويسبب نوعاً من الالتباس، وللحفــاظ على حجم العينة الصحيح يمكننا ضرب معكوس الاحتمــالات فــي نــسبة الأشخاص الحقيقيين إلى الوهميين 940/503 فإن هذه الخطوة سوف تقــوم بإنشاء متغير جديد باسم *adultwt* وهو يحتــوي علــى الأوزان الاحتماليــة لتصحيح تحيز العينة المعروف، مع الحفاظ على حجــم العينــة الأصــلي. الأوزان تساوي 0.535 (شخص بالغ واحد بكل منزل)، 1.070 (شخــصان بالغان بكل منزل)، 1.605 (ثلاثة أشخاص أو أكثر بالغين بكل منزل)، القيم المفقودة سوف تأخذ الوزن المحايد وهو 1، ونسبة هذه الأوزان تظل 3:2:1

```
.generate adultwt = adults*(503/940)
.replace adultwt = 1 if missing(adultwt)
.tab adults, summ(adultwt) miss
```

# adults in household	Summary of adultwt Mean	Std. Dev.	Freq.
1	.53510636	0	148
2	1.0702127	0	273
3+	1.6053191	0	82
DK/NA	1	0	13
Total	.99999997	.35080553	516

إذا كان هذا التعديل مرغوباً فيه، فيمكننا استخدام الأمر svyset مع بيانات المتغير *adultwt* كأوزان احتمالية.

```
.svyset _n [pw = adultwt]

      pweight: adultwt
          VCE: linearized
  Single unit: missing
     Strata 1: <one>
         SU 1: <observations>
        FPC 1: <zero>
```

```
.svy: tab adults, percent
(running tabulate on estimation sample)

Number of strata   =          1            Number of obs     =        503
Number of PSUs     =        503            Population size   = 502.99998
                                           Design df         =        502
```

# adults in household	percentages
1	15.74
2	58.09
3+	26.17
Total	100

Key: percentages = cell percentages

النسب الموزونة (مثل 16% من شخص واحد بالمنزل بدلاً من 29% كما في البيانات الخام) يعطي صورة أكثر واقعية.

الأوزان المرجحة الطبقية اللاحقة : Poststratification Weights

الجزء السابق في هذا الفصل، قدّم مثالاً على الأوزان المرجحة بناءً على تصميم المعاينة، والتي كانت معروفة قبل تصميم عملية جمع البيانات. النوع الثاني من الأوزان المرجحة يمكن تعريفه بعد جمع البيانات. فبالرغم من توخي الدقة عند جمع البيانات، فإنه من الممكن أن البيانات لا تمثل بعض خصائص المجتمع، فمثلاً قد يكون هناك اختلاف واضح بين توزيع العمر أو

الجنس في العينة عند توزيع المجتمع المستهدف، مما يجعل النتائج عرضة للتساؤل، فالتوزيع الطبقي اللاحق، يشير إلى الأوزان الاحتمالية المحسوبة حتى تكون نسب مجموعات معينة أو طبقات في العينة قريبة بدرجة معقولة لما هو موجود في المجتمع.

ففي عينة مركز جرانيت لاستطلاع الرأي، كانت نسبة الإناث 54.65%، ولكن حسب تعداد السكان لسنة 2010 نسبة الإناث البالغات بولاية هامبشير 51.6% فقط، وإذا أظهرت نتائج الدراسة الاستقصائية بأن نسبة الإناث بالمجتمع 54.65% فهذا يكون جنوحاً كبيرًا عن الواقع. بالإضافة إلى ذلك، فقد نحصل على نتائج خاطئة حول العناصر الأخرى المرتبطة بالجنس مثل الانتخاب. هذا التحيز الواضح في عدد الردود يمكن أن يؤثر على قدرتنا في الحصول على استدلالات صحيحة في مجتمعات أكبر.

```
.tab sex
```

Gender	Freq.	Percent	Cum.
Male	234	45.35	45.35
Female	282	54.65	100.00
Total	516	100.00	

هناك العديد من الطرق للوصول إلى التقسيم الطبقي اللاحق (الطريقة البديلة للأسلوب اليدوي الموضح أدناه هو استخدام الأمر **svyset** والذي يوفر خيار **poststrata** الذي تم شرحه بالتفصيل بدليل المستخدم *Survey Reference Manual*) إذا كنا نعرف نسب المجتمع الصحيحة للمتغيرات الرئيسة – كما فعلنا سابقاً بخصوص الجنس بالمثال أعلاه – فإن أوزان تصحيح تحيز الإجابات يمكنه حسابها من خلال قسمة نسب المجتمع على نسب العينة، فمثلاً متغير الجنس *sex* تم ترميزه بحيث يساوي 0 للذكور، وهم يمثلون 48.4% من البالغين في المجتمع بولاية هامبشير، ولكن نسبتهم في العينة هي 45.35% فقط وليس هناك أي قيم مفقودة لمتغير الجنس *sex* في بيانات العينة.

يمكننا حساب الأوزان حيث إنها أعلى من 1 بقليـل 48.4 ÷ 45.35 = 1.067 للذكور، أما للإناث فهي أقل من 1 بقليل 51.6÷54.65=0.944

```
.generate sexwt = 48.4/45.35 if sex==0
.replace sexwt = 51.6/54.65 if sex==1
.tab sex, summ(sexwt)
```

Gender	Summary of sexwt Mean	Std. Dev.	Freq.
Male	1.0672547	0	234
Female	.94419032	0	282
Total	.99999857	.06132481	516

إذا استخدمنا الأمر **svyset** مع بيانات المتغير *sexwt* كوزن احتمالي، فإن الخيار **svy:tab** يقوم بإنتاج جدول مرجح يعرض النسب الحقيقية للذكور وهي 48.4% وللإناث 51.6%. بعد حساب الأوزان الطبقية اللاحقة من الأفضل فحص ما إذا كانت الأوامر قد قامت بوظيفتها كما ينبغي أم لا؟

```
.svyset [pw = sexwt]

      pweight: sexwt
          VCE: linearized
  Single unit: missing
     Strata 1: <one>
         SU 1: <observations>
        FPC 1: <zero>
```

```
.svy: tab sex, percent
(running tabulate on estimation sample)

Number of strata   =         1          Number of obs      =        516
Number of PSUs     =       516          Population size    = 515.99926
                                        Design df          =        515
```

Gender	percentages
Male	48.4
Female	51.6
Total	100

Key: percentages = cell percentages

من الممكن حساب أوزان طبقية لاحقة أخرى باتباع طريقة مشابهة، فمثلاً إذا افترضنا أنه في دراسة أخرى نريد تقدير توزيع العمر والعرق والجنس في المجتمع، فيمكن أن يتم ذلك باتباع الخطوات التالية:

1- إنشاء جدول لنسب العمر والعرق والجنس من تعداد سكاني أو بيانات أخرى مثل عدد البالغين القاطنين بالمنزل عن المجتمع المستهدف، فإذا قمنا باستخدام خمس مجموعات للعمر (18-29، 30-39.. الخ) ومجموعتين للعرق (أبيض، غير أبيض) فإن النتائج سوف تكون في 20 رقم (5×2) مثل نسبة البالغين البيض الذكور في المجتمع في مجموعة 18-29 سنة، أو نسبة الإناث البيض في مجموعة 18-29 سنة وهكذا.

2- الحصول على جدول مشابه للجدول السابق يوضح نسب العمر والعرق والجنس من المثال، فمثلاً عند إنشاء متغير جديد باسم *ARS* يجمع بين العمر والعرق والجنس وإنشاء جدول له:

```
.egen ARS = group(agegroup race sex),
lname(ars)
.tab ARS
```

3- تعريف مجموعة أوزان جديدة باستخدام الأمر **generate ... if**. فعلى سبيل المثال، بافتراض أننا نعلم بأن 8.6% من عدد السكان البالغين في منطقة هذه الدراسة هم من الذكور البيض في مجموعة 18-29 سنة، وأن نسبة 8.2% هم من الإناث البيض في نفس المجموعة، ولكن في العينة غير الموزونة نرى أن نسبة 2.6% من الذكور البيض في مجموعة 18-29 سنة وأن نسبة الإناث البيض 5.1% مما يعني أن الذكور يشكلون النسبة الأكبر في عدد البالغين، وهذا لا يمثل النسب الحقيقية بالمجتمع، يمكننا إنشاء متغير موزون جديد للعمر والعرق والجنس يسمى *ARSwt* يساوي 1 (وهو الوزن المحايد)، وإذا كنا لا نعرف عمر وعرق وجنس المشارك في الدراسة، وغير ذلك، فإنه يساوي نسبة المجتمع. وقسمة نسبة العينة بالنسبة لمجموعة العمر والعرق والجنس كما يلي:

```
.generate ARSwt = 1 if ARS>= .
```

```
.label variable ARSwt "Age-race-sex weights"
.replace ARSwt = 8.6/2.6 if ARS == 1
.replace ARSwt = 8.2/5.1 if ARS == 2
```

تصحيح التقسيم الطبقي اللاحق طريقة مفيدة عند العمل مع دراسات استقصائية مصممة تصميماً جيدًا، ويجب ألا يُفهم هذا على أنه علاج للأخطاء التي تحدث في المعاينة بالصدفة. مثل هذه التصحيحات يمكن تطبيقها بشكل كبير على دراسات استطلاع رأي الناخبين، والدراسات الاستقصائية بالعلوم الاجتماعية التي تتطلب مجهودًا كبيرًا للحصول على عينة ممثلة للمجتمع. وهناك بعض الدراسات التي تبحث عن دليل مستقل مثل نتائج الانتخاب أو التي يتم القيام بها من جديد من قبل باحثين آخرين، فهذه تعتبر اختبارات حقيقية عن مدى نجاح التصحيحات.

بيانات دراسة واحدة قد تتضمن وزن متغيرات تم حسابها من أكثر من مصدر، مثل تصميم الأوزان، والأوزان الطبقية اللاحقة، ولدمج هذه المتغيرات في متغير مرجح عام واحد نقوم بضرب ثم نقوم بعمل تصحيح حتى يكون المجموع النهائي للأوزان يساوي حجم العينة. فعند وضع الأمر **quietly** قبل الأمر **summarize**، فإننا نطلب من ستاتا حساب الإحصائيات المختصرة، ولكن لا تعرض النتائج توفيراً للوقت. واستخدام الأمر **quietly** مع الأوامر الأخرى، فإنه سوف يقوم بنفس المهمة.

```
.generate finalwt = adultwt*ARSwt
.replace finalwt = 1 if finalwt>= .
.quietly summarize finalwt
.replace finalwt = finalwt*(r(N)/r(sum))
```

يمكن وجود أي عدد من المتغيرات المرجحة في نفس مجموعة البيانات، واستخدام الأمر **svyset** باستمرار لاختيار أي متغيرات لإجراء تحليل معين، الأوزان تتأثر في التحليل عند تطبيقها فقط باستخدام الأمر **svy** أو تحديد أوامر ترجيح أخرى. خلال بقية هذا الفصل، سوف نرجع إلى بيانات

استطلاع الرأي لمركز جرانيت، ونقوم بترجيحها بواسطة المتغير *censuswt* وهو متغير تم حسابه بواسطة مركز الاستطلاع بجامعة نيوهامبشير بالولايات المتحدة لتوحيد تصميم الأوزان (لعدد البالغين، وعدد خطوط الهاتف)، مع تصنيف طبقي لاحق (للجنس والمنطقة بولاية هامبشير).

```
.svyset _n [pw = censuswt]

      pweight: censuswt
          VCE: linearized
  Single unit: missing
     Strata 1: <one>
         SU 1: <observations>
        FPC 1: <zero>
```

الرسومات البيانية والجداول الموزونة للدراسات الاستقصائية :

Survey-Weighted Tables and Graphs

استطلاع يونيو 2011 لمركز جرانيت، يتضمن ستة أسئلة تتعلق بالاحتباس الحراري، أو التغير المناخي. مجموعة من هذه الأسئلة كانت حقيقية، ولكن واحدًا منها (*warmop*) كان عما هو اعتقادك الشخصي؟

أي من العبارات الثلاث التالية تعتقد أنها صحيحة؟

- التغير المناخي يحدث الآن وسببه الرئيسي الأنشطة البشرية.
- التغير المناخي يحدث الآن ولكن سببه الرئيسي القوى الطبيعية.
- التغير المناخي لا يحدث الآن.

الباحث قام بتغيير ترتيب خيارات الإجابات لتفادي احتمالية التحيّز. نحو 55% وافقوا بأن التغير المناخي يحدث الآن وسببه الرئيس الأنشطة البشرية، وآخرون يعتقدون بأن التغيرات كانت لأسباب طبيعية (35%)، قليلون يعتقدون بأن التغير المناخي لا يحدث الآن (3%).

```
.svy: tab warmop, percent ci
```

```
(running tabulate on estimation sample)

Number of strata   =         1          Number of obs      =       516
Number of PSUs     =       516          Population size    = 515.57392
                                        Design df          =       515
```

Personal belief about climate change	percentages	lb	ub
DK/NA	7.443	5.252	10.45
Not now	3.05	1.904	4.853
Now/natu	34.62	30.2	39.32
Now/huma	54.89	50.11	59.58
Total	100		

```
Key:  percentages  =  cell percentages
      lb           =  lower 95% confidence bounds for cell percentages
      ub           =  upper 95% confidence bounds for cell percentages
```

الأمر **svy:tab** : يُطبق الأوزان بناءً على معايير تم تحديدها سابقاً بواسطة الأمر **svyset**، والخيار **ci** يحدد فترات الثقة للنسبة الموزونة ويعرضها كحدود سفلى وعليا (الحد الأدنى lb والحد الأعلى ub)، وبناء على هذه العينة فإننا على درجة ثقة 95% بأنه ما بين 50.11% و 59.58% من البالغين بولاية هامبشير يعتقدون بأن الأنشطة البشرية تقوم بتغيير المناخ.

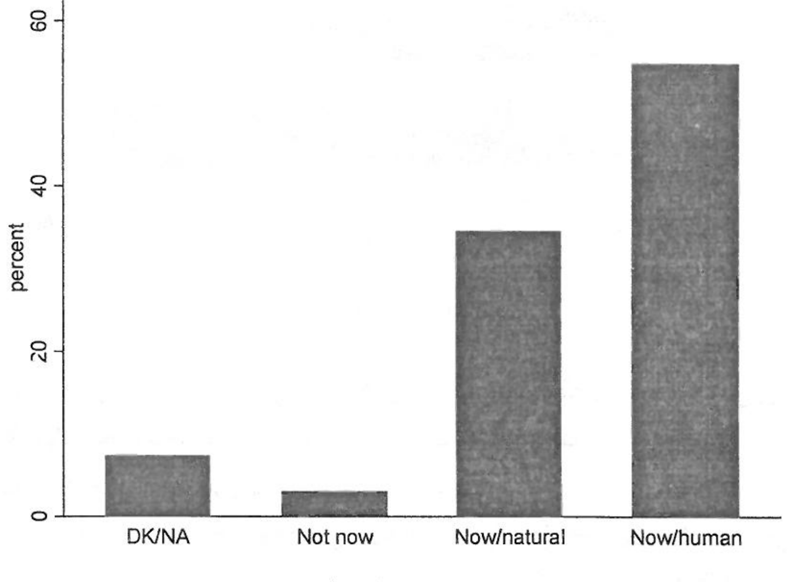
أنواع الرسومات البيانية الأصلية ببرنامج ستاتا ليست مناسبة لعرض توزيعات المتغيرات الطبقية مثل المعروضة بالجدول أعلاه، لحسن الحظ هناك برنامج يمكن للمستخدم كتابته اسمه **catplot** – تم شرحها في مجلة ستاتا *Stata Journal* للكاتب كوكس (Cox 2004b) – يقوم بهذه الوظيفة بطريقة جيدة، ويمكنك الحصول على الملفات التنفيذية do-files لهذا البرنامج من الإنترنت وذلك بطباعة الأمر

```
.findit catplot
```

وتابع الروابط لتحميل وتثبيت هذه البرامج على جهاز الكمبيوتر لديك، (الأمر **findit** يعمل مع مئات من البرامج المكتوبة بواسطة المستخدمين) وعند إنهاء التثبيت قم بطباعة الأمر **help catplot** لعرض تركيبة الأمر وخياراته، الشكل (1.4) يحتوي على رسم بياني عمودي للمتغير *warmop*. وبالرغم من

أن الأمر catplot لا يقبل وضع الأعمدة البيانية في وضع عمودي ويقبل الأعمدة الأفقية فقط، فإن إضافة الخيار [aweights = *censuswt*] في هذا الأمر يجعل له نفس التأثير المرئي لكي تكون الأعمدة في وضع عمودي وليس أفقياً موضحةً النسب المئوية svy: tab.

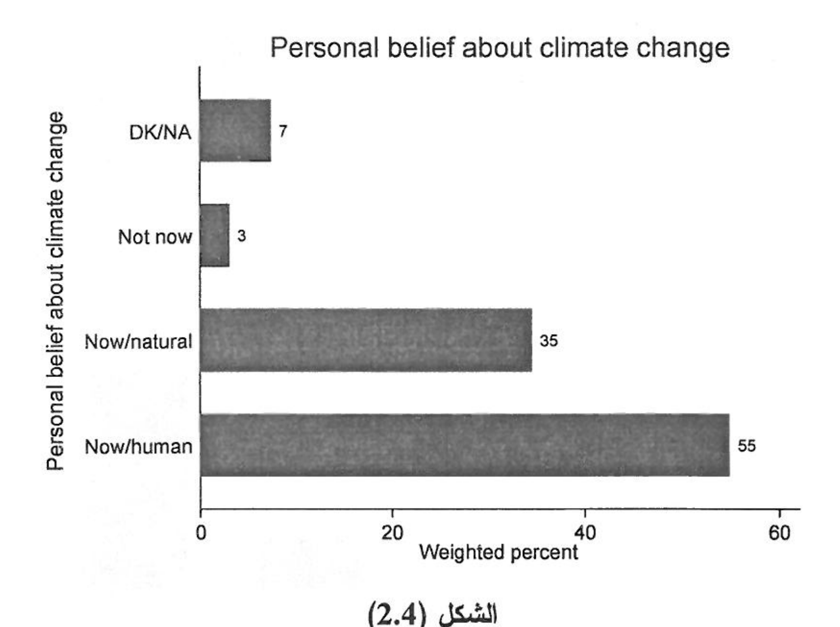
```
.catplot bar warmop [aweight = censuswt], percent
```



الشكل (1.4)

الرسم البياني العمودي مع توصيفات القيم في العادة أسهل للقارئ من الوضع الأفقي بتنسيق (hbar) وخصوصاً عندما يكون لدينا العديد من الأعمدة. الشكل (2.4) يعرض تنسيقاً أفقياً يتضمن عنواناً وتوصيفات للمحاور بشكل مناسب للنشر في التقارير أو لعرض نتائج الدراسة الاستقصائية، يمكننا توصيف الأعمدة حتى يمكن قراءة النسب المرجحة مباشرة من الرسم البياني، وهناك نفس العدد تم إيجاده بواسطة الأمر أعلاه svy: tab.

```
.catplot hbar warmop [aweight= censuswt], percent
blabel(bar, format(%3.0f)) ytitle("Weighted
percent")
title("Personal belief about climate change")
```



الشكل (2.4)

كيف يمكن للتغير المناخي أن يكون متعلقاً بـالمتغيرات الأخـرى فـي الدراسة الاستقصائية، مثل مستوى تعليم المشارك بالدراسة (*educ*)؟ يمكننـا إجابة مثل هذه الأسئلة من خلال الجداول الثنائية.

```
.svy: tab warmopeduc, col percent
(running tabulate on estimation sample)

Number of strata   =         1          Number of obs      =       511
Number of PSUs     =       511          Population size    = 510.02315
                                        Design df          =       510

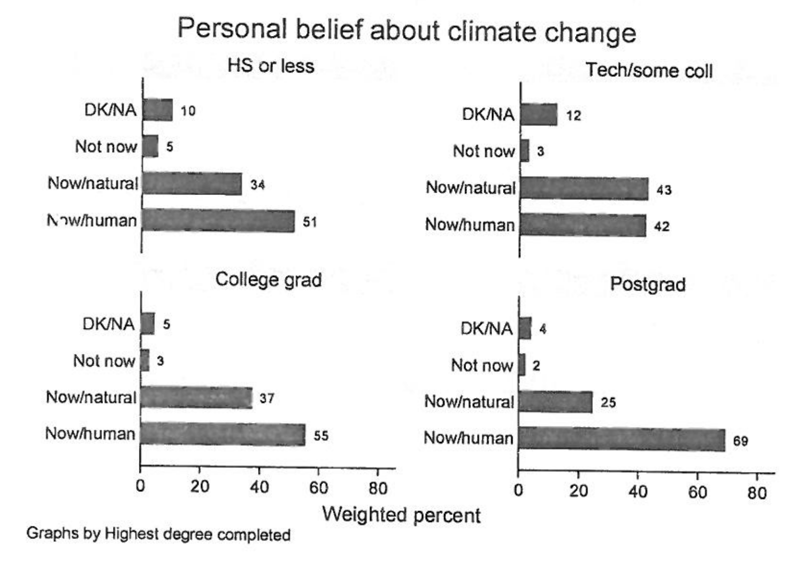
----------------------------------------------------------------------
Personal |
belief   |
about    |
climate  |           Highest degree completed
change   |  HS or le  Tech/som   College  Postgrad     Total
---------+------------------------------------------------------------
   DK/NA |     9.946     12.27     4.603     4.041     7.524
 Not now |     5.154     2.694     2.694     1.991     3.083
Now/natu |     33.55     42.81     37.29     24.79     34.71
Now/huma |     51.35     42.23     55.41     69.18     54.68
         |
   Total |       100       100       100       100       100
----------------------------------------------------------------------
  Key:  column percentages

  Pearson:
    Uncorrected   chi2(9)          =   25.1986
    Design-based  F(8.81, 4495.16)=     2.4226     P = 0.0102
```

في المثال أعلاه، قمنا بإنشاء عمود للنسب المئوية، لأن عمود المتغير (*educ*) يشكل متغيرًا مستقلاً في هذا التحليل، التحليل يوضح أن نحو 69% هم من حملة شهادة دراسات عليا Postgrad، 55% خريجو كليات College، 42% جامعي لم يكمل دراسته أو درس في معهد تقني Tech/some، جميعهم يعتقدون بأن التغير المناخي يحدث الآن وسببه الرئيس أنشطة إنسانية، والتصميم المرجح لاختبار *F* في الجدول أعلاه، يؤكد على أن العلاقة بين التغير المناخي، ومستوى التعليم ذو معنوية إحصائية (p=0.102).

الشكل (3.4) يعرض رسماً بيانياً من نوع **catplot** للمتغير *warmop* موضحاً الردود لكل مستوى تعليمي للمشاركين، معطياً نسباً مئوية مرجحة. الخيار **percent(educ)** يقوم بتصنيف النسب المئوية حسب فئات مستوى التعليم *educ* والخيار **title()** تم استخدامه كخيار فرعي مع المحدد **by()** ولكي نحصل على عنوان واحد للرسم البياني ككل، يمكنك اختبار ذلك لترى ما سيحدث بدون استخدام هذه الخيارات.

```
.catplot hbar warmop [aweight = censuswt],
percent(educ)
by(educ,title("Personal belief about climate
change"))
blabel(bar, format(%3.0f)) ytitle("Weighted
percent")
```



الشكل (3.4)

مخططات الأعمدة البيانية للمقارنات المتعددة :

Bar Charts for Multiple Comparisons

الرسم البياني للأعمدة **catplot** في الشكل (3.4) يوضح العلاقة بين متغيرين، كل منهما يحتوي على أربع فئات. وإذا كان لدينا أكثر من متغيرين أو عدد كبير من الفئات، فإن استخدام الأمر **catplot** يصبح معقدًا، البديل الأكثر وضوحاً لإجراء المقارنات المتعددة للمتغيرات هو استخدام الأمر **hbar** لإنشاء الأعمدة الأفقية.

الشكل (13.3) في الفصل السابق، تتبع التغيرات في جليد القطب الشمالي في نهاية فصل الصيف للفترة من 1979-2011، الانخفاض الكبير في نسبة الجليد جذب انتباه الكثير من العلماء وتم ملاحظته من قبل العامة في النشرات ووسائل الإعلام المختلفة، وقد قام مركز جرانيت لاستطلاع الرأي بتضمين سؤال (*warmice*) تم صياغته بعناية لاختبار مدى معرفة الناس عن هذه المشكلة، مع إضافة هذا السؤال تم تغيير ترتيب الإجابات لتجنب التحيز، الأغلبية الساحقة (71%) تعلم عن انخفاض الجليد بالقطب الشمالي.

أي من العبارات الثلاث التالية تعتقد أنها الأكثر دقة؟

خلال السنوات القليلة الماضية، الجليد في القطب الشمالي في نهاية الصيف:

- يغطي منطقة أقل عن التي كان يغطيها منذ 30 سنة مضت.
- انخفض ولكن عاد لنفس المنطقة تقريباً التي كان يغطيها منذ 30 سنة مضت.
- يغطي منطقة أكثر من التي كان يغطيها منذ 30 سنة مضت.

```
.svy: tab warmice, percent ci
```

```
(running tabulate on estimation sample)

Number of strata   =         1          Number of obs      =       516
Number of PSUs     =       516          Population size    = 515.57392
                                        Design df          =       515
```

Arctic ice vs. 30 years ago	percentages	lb	ub
Less	70.91	66.35	75.08
Recovere	10.43	7.784	13.83
More	6.916	4.841	9.789
DK/NA	11.75	8.991	15.21
Total	100		

```
Key:  percentages  =  cell percentages
      lb           =  lower 95% confidence bounds for cell percentages
      ub           =  upper 95% confidence bounds for cell percentages
```

سؤال *warmice*: يسمح بأربعة خيارات تتضمن "لا أعرف" "don't know" أو لا إجابة، وخدمة لبعض الأغراض، فإنه من المفيد إنشاء متغير جديد ذي تفرعين يشير إلى ما إذا كان المشاركون قد أجابوا عن السؤال بشكل صحيح، المتغير *warmiceQ* يساوي 1 للإجابات التي قالت بأن الجليد انخفض، 0 لجميع الإجابات الأخرى، نحو 71% أجابوا بشكل صحيح، وكانت إجاباتهم أن الجليد في نهاية الصيف يغطي منطقة أقل من تلك التي كان يغطيها منذ 30 سنة مضت.

```
.gen warmiceQ = 0
.replace warmiceQ = 1 if warmice==1
.label variable warmiceQ "Know Arctic ice area
 declined"
.svy: tab warmiceQ, percent ci
```

```
(running tabulate on estimation sample)

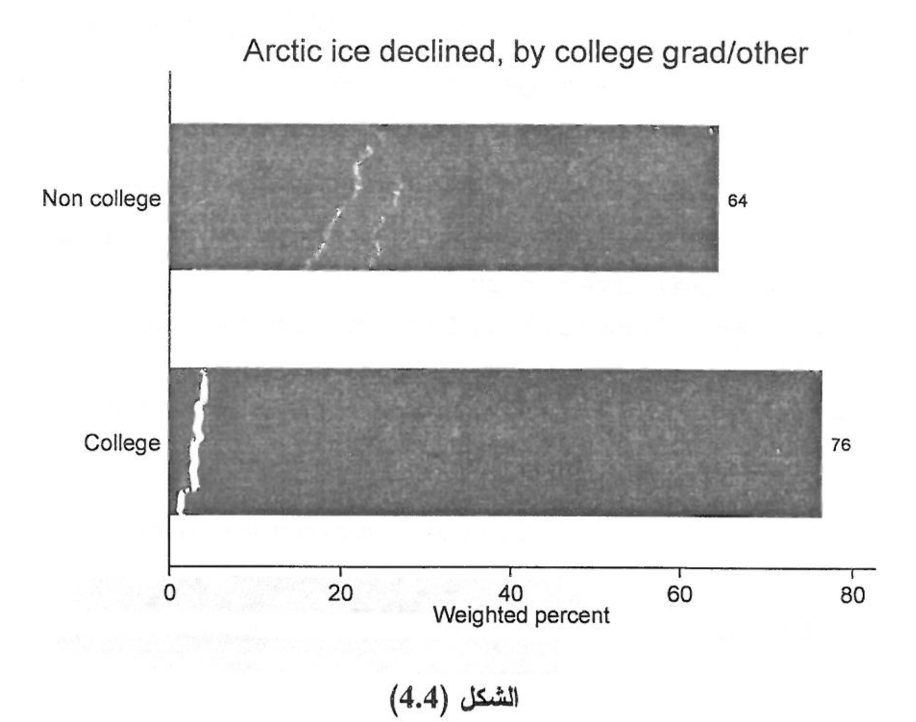
Number of strata   =          1                 Number of obs      =         516
Number of PSUs     =        516                 Population size    = 515.57392
                                                Design df          =         515
```

Know Arctic ice area declined	percentages	lb	ub
0	29.09	24.92	33.65
1	70.91	66.35	75.08
Total	100		

```
Key:  percentages  =  cell percentages
      lb           =  lower 95% confidence bounds for cell percentages
      ub           =  upper 95% confidence bounds for cell percentages
```

المتوسط {0,1} لمتغير معين مثل *warmiceQ* يساوي نسبة من قيم 1 الصحيح، فمثلاً المتغيرات الوهمية {0,1} لها العديد من الاستخدامات في النماذج الإحصائية. فعند إنشاء الرسومات البيانية فقد يتم إعادة قياس المتغير *warmiceQ* لتكون قيمه 0 أو 100، فالمتغير الذي متوسطه {0,100} يساوي نسباً مئوية. وفي المثال أعلاه فهو يساوي النسب المئوية للإجابات الصحيحة حول الجليد في القطب الشمالي، وعند تطبيق الأمر **graph hbar** على تقسيم ثنائي {0,100} فيمكننا مقارنة أشكال العديد من النسب المئوية، الشكل (4.4) يعرض النسب المرجحة للإجابات الصحيحة لخريجي الجامعات وخريجي المؤسسات الأخرى (المتغير *college*).

```
.gen warmiceQ100 = warmiceQ*100
.graph hbar (mean) warmiceQ100 [aw = censuswt],
 over(college)
blabel(bar, format(%3.0f)) ytitle("Weighted
 percent")
title("Arctic ice declined, by college
 grad/other")
```



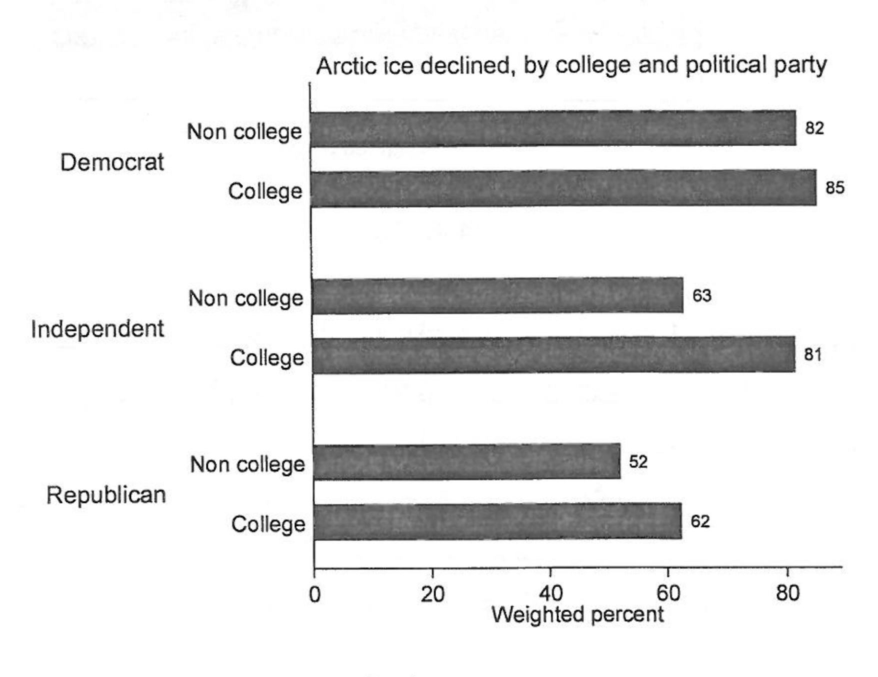
الشكل (4.4)

لاحظ أن الرسم البياني للأعمدة الأفقية (**graph hbar** وأيضاً في **graph hbox**، وفي بعض الأشكال البيانية الأخرى التي تم تغيير اتجاهها) المحور الأفقي يسمى محور y ، وهذا مخالف لما هو متعارف عليه، لذا فإن **ytitle("Weighted percent")** يحدد عنواناً للمحور الأفقي وهو يظهر في أسفل الرسم. فبرنامج ستاتا لا يتعرّف على محور x في مثل هذه الأشكال بالرغم من أن الخيار **l1title( )** يمكنه تحديد عنوان يظهر في الجانب الأيسر للشكل البياني.

الشكل (4.4) يقارن بين نسبتين فقط وهي نسبة 76% التي تمثل نسبة خريجي الجامعات ونسبة 65% التي تمثل خريجي المؤسسات الأخرى. لا أعتقد أن هناك شخصاً يريد إنشاء مثل هذا الرسم البياني لإجراء مثل هذه المقارنة البسيطة، ولكن طريقة رسم الأعمدة يمكن تحسينها لإجراء مقارنات أكثر تعقيداً. فالدراسات الاستقصائية السابقة وجدت أن هناك اختلافات

متحيزة في العديد من الأسئلة التي تتعلق بالتغير المناخي، وقد توجد هذه المشكلة في السؤال السابق المتعلق بالجليد في القطب الشمالي. والشكل (5.4) يعرض أعمدة بيانية لثلاثة متغيرات تعطي نسبة الإجابات الصحيحة، وتم تقسيم هذه النسب لتمثل التعليم الجامعي والانتماء السياسي (متغير *party*).

```
.graph hbar (mean) warmiceQ100 [aw = censuswt],
over(college) over(party)
blabel(bar, format(%3.0f)) ytitle("Weighted
percent")
title("Arctic ice declined, by college and
political party", size(medium))
```



الشكل (5.4)

لكل مجموعة أحزاب بالمتغير *party*، يمكننا أن نرى تأثير التعليم الجامعي ولكل مستوى تعليمي، يمكننا أيضاً أن نرى الاختلافات المتحيّزة.

الشكل (5.4) لا يستطيع أن يوضح لنـــا أن الاختلافـــات هـــي اختلافـــات إحصائية ذات معنوية إحصائية، فالإجابة عن ذلك تتطلب استخدام أدوات النماذج الإحصائية التي سوف يتم تقديمها في الفصل (9)، كما ســـبق وأن رأينـــا، فـــإن نموذج الانحدار اللوغاريتمي المرجح يؤكد بأن كلاً من المتغيـــر *college* والـــذي يمثل التعليم الجامعي، والمتغير *party* والذي يمثل الانتماء السياسي لها تـــأثيرات إحصائية ذات معنوية، وهذا التأثير موجب في حالة *college* (0 = غير الجامعي non-college، 1 = الخريج الجامعي college grad) وسالب في حالـــة *party* (1 = الديمقراطي Democrat، 2= للمستقل Independent، 3 = للجمهوري Republican)، والانتماء السياسي عامل تنبؤي أقوى من *college*.

```
.svy: logit warmiceQ college party
(running logit on estimation sample)

Survey: Logistic regression

Number of strata   =        1          Number of obs      =        501
Number of PSUs     =      501          Population size    =  500.96122
                                       Design df          =        500
                                       F(   2,     499)   =      16.07
                                       Prob > F           =     0.0000
```

warmiceQ	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
college	.4634607	.2264922	2.05	0.041	.018467	.9084544
party	-.6669759	.1268445	-5.26	0.000	-.9161897	-.4177621
_cons	2.058491	.3162993	6.51	0.000	1.437052	2.679931

سوف نعود لهذا المثال في الفصل 9، حيث سوف يتم تطبيـــق طريقـــة إحصائية (اللوغاريتم المتعدد) ويمكن إدراجها في نموذج تنبؤ لكل إجابة عن السؤال الخاص بمناطق الجليد.



الفصل الخامس

الملخصات الإحصائية والجداول

Summary Statistics and Tables

الأمر summarize: يقوم بإنشاء إحصائيات وصفية مختصرة، مثل الوسيط والمتوسط والانحراف المعياري للمتغيرات. وهناك طرق أخرى لإنشاء الملخصات الإحصائية للمتغيرات الأخرى، وذلك باستخدام الأمر **tabstat**. بالنسبة للمتغيرات الطبقية والترتيبية، فإن الأمر **tabulate** يقوم بإنشاء جدول للتوزيع التكراري، والجداول التقاطعية، ومجموعة من الاختبارات، وقياسات للعلاقات، كما يمكنه أيضاً إنشاء جداول أحادية أو ثنائية للمتوسطات والانحرافات المعيارية في شكل فئات للمتغيرات الأخرى. الأمر العام لإنشاء الجداول "**table**" هو مقدمة لستة أنواع من الجداول، تحتوي خانات هذه الجداول على إحصائيات مثل التكرارات والمجاميع والمتوسطات والوسيط، وأخيراً سوف نراجع إجراءات المتغير الواحد، وهي تتضمن اختبارات الاعتدال والتحويلات وعرض تحليل البيانات الاستطلاعية (EDA). وأغلب التحليلات التي يغطيها هذا الفصل، يمكن القيام بها من خلال الأوامر أو من خلال القوائم وذلك باختيار Statistics > Summaries, tables & tests.

بالإضافة إلى هذه التحليلات العامة، يقوم ستاتا بإنشاء العديد من الجداول التي لها أهمية خاصة لدى علماء الأوبئة، وقد قام الكاتب Selvin (2004) بالتطرّق لهذا الموضوع بالتفصيل.

أمثلة عن الأوامر : Example Commands

```
.summarize y1 y2 y3
```

يقوم هذا الأمر بحساب ملخص للإحصائيات (المتوسطات والانحرافات المعيارية وأعلى وأقل قيمة وعدد المشاهدات) للمتغيرات المدرجة بالأمر.

```
.summarize y1 y2 y3, detial
```

يقوم هذا الأمر بإنشاء ملخصات إحصائية أكثر تفصيلاً تتضمن نسباً مئوية والوسيط والمتوسط الحسابي والانحراف المعياري والتباين والالتواء والتفرطح.

```
.summarize y1 if x1>3 & !missing(x2)
```

يقوم بتحديد ملخص لإحصائيات المتغير *y1* مستخدماً المشاهدات التي قيمتها أكبر من 3 للمتغير *x1* والمشاهدات الموجودة (غير المفقودة) للمتغير *x2*

```
.summarize y1 [fweight = w], detail
```

يقوم بحساب ملخصات إحصائية أكثر تفصيلاً للمتغير *y1* باستخدام الأوزان التكرارية في المتغير *w*.

```
.tabstat y1, stats(mean sd skewness kurtosis n)
```

يقوم بحساب الإحصائيات التي تم تحديدها بين الأقواس للمتغير *y1*.

```
.tabstat y1, stats(min p5 p25 p50 p95 max) by(x1)
```

حساب ملخصات إحصائية محددة (أقل قيمة، المئين 5، والمئين 25 وهكذا) للمتغيرات *y1* مستخدماً فئات المتغير *x1*.

```
.tabulate x1
```

يعرض جدولاً للتوزيع التكراري لكل القيم الموجودة للمتغير *x1*.

```
.tabulate x1, sort miss
```

يعرض جدول توزيع تكراري للمتغير *x1* يتضمن القيم المفقودة، ويتم ترتيب الصفوف (القيم) من أعلى تكرار إلى أقل تكرار.

```
.tab1 x1 x2 x3 x4
```

يعرض سلسلة من جداول التوزيع التكراري بحيث يتم إنشاء جدول تكراري لكل متغير.

```
.tabulate x1 x2
```

يعرض جدولاً تقاطعياً لمتغيرين، بحيث إن المتغير *x1* يكون في صفوف الجدول، والمتغير *x2* يكون في الأعمدة.

```
.tabulate x1 x2, chi2 nof column
```

يقوم بإنشاء جدول تقاطعي، وإجراء اختبار بيرسون χ^2 للاستقلال، ولايعرض خلايا التكرارات ولكن يعطي عمودًا للنسب في كل خلية.

```
.tabulate x1 x2, missing row all
```

يقوم بإنشاء جدول تقاطعي يتضمن القيم المفقودة في الجدول وحسابات النسب المئوية، كما يحسب كل إحصائيات المتغيرات (بيرسون واحتمال χ^2، وقيمة V لكرامر Cramér's V، وقيمة جاما لجودمان وكروسكال Goodman and Kruskal's gamma، قيمة τ_b لكندال Kendall).

```
.tab2 x1 x2 x4 x4
```

يقوم بإنشاء جدول تقاطعي ثنائي للمتغيرات المدرجة بالأمر.

```
.tabulate x1, summ(y)
```

يقوم الأمر بإنشاء جدول أحادي يعرض المتوسط والانحراف المعياري والتكرارات لقيم المتغير *y* لكل فئة من فئات المتغير *x1*.

```
.tabulate x1 x2, summ(y) means
```

يقوم بإنشاء جدول تقاطعي ثنائي يعرض متوسط المتغير *y* عند كل مرافق من قيم المتغيرات *x1* والمتغير *x2*.

```
.by x3, sort: tabulate x1 x2, exact
```

يقوم بإنشاء جدول تقاطعي ثلاثي مع جداول فرعية للمتغير *x1* (الصف) والمتغير *x2* (عمود) عند قيمة من قيم المتغير *x3*، كما يقوم بحساب اختبار الدقة لفيشر Fisher's exact لكل جدول فرعي.

الخيار by *varname*, sort: يعمل كمقدمة لأغلب أوامر ستاتا عندما يكون له معنى، والخيار **sort** غير ضروري إذا كانت البيانات مرتبة في المتغير *varname*.

```
.table y x2 x3, by(x4 x5) contents(freq)
```

يقوم بإنشاء جدول تقاطعي خماسي للمتغير *y* (صف) في المتغير *x2* (عمود) في المتغير *x3* (عمود فرعي) في المتغير *x4* (صف فرعي 1) في

المتغير $x5$ (صف فرعـــي 2) وجميع الخلايـــا ســـوف تحتـــوي علـــى تكرارات.

```
.table x1 x2, contents(mean y1 median y2)
```

يقوم هذا الأمر بإنشاء جدول تقاطعي ثنائي للمتغيـــر $x1$ (صـــف) فـــي المتغير $x2$ (عمود)، والخلايا سوف تتضمن المتوسط للمتغير $y1$ والوســـيط للمتغير $y2$.

```
.svy: tab y, percent ci
```

يقوم هذا الأمر باستخدام البيانات المرجحة للدراسات الاستقصائية (التي تم تحديدها بالأمر **svyset**) وإنشاء جدول ثنائي للنسبة المئوية للمتغير y مـــع فترة ثقة 95%. وللحصول على مزيد من المعلومات عن خيارات جـــداول بيانات الدراسات الاستقصائية، قم بطباعة الأمر **help svy tab**. كما أن الفصل 14 سيشرح بيانات الدراسات الاستقصائية وكيفية تحليلها.

```
.svy: tab y x, column percent
```

يقوم هذا الأمر باستخدام البيانات المرجحة للدراسة الاستقصائية، وإنشاء جدول ثنائي مع صف للمتغير y وعمـــود للمتغير x وعرض نتيجة اختبـــار الاستقلاليـــة المعدل χ^2، وســـوف تحتـــوي الخلايـــا علـــى نـــسب مئويـــة مرجحة.

الملخصات الإحصائية لقياس المتغيرات :

Summary Statistics for Measurement Variables

الملف *electricity.dta* يتضمن بيانات ومعلومات عن استهلاك الكهربـــاء في الولايات المتحدة. وتم الحصول على هذه البيانات من مفوضـــية الطاقـــة بكاليفورنيا (2012).

```
.use C:\data\electricity.dta, clear
.describe
```

```
Contains data from C:\data\electricity.dta
  obs:            51                         US states
                                               electricity use
                                               2010 (CA Energy
                                               Commission)
 vars:             7                         2 Jul 2012 06:11
 size:         1,734
------------------------------------------------------------------
              storage  display    value
variable name   type   format     label      variable label
------------------------------------------------------------------
state           str20  %20s                  State
stateab         str2   %9s                   State
                                          (abbreviation)
region4         byte   %9.0g      reg4       Census Region (4)
region9         byte   %12.0g     reg9       Census Division
                                          (9)
pop             long   %8.0g                 Population, 1000s
electric        long   %8.0g                 Electricity use,
                                          millions of kWh
elcap           int    %8.0g                 Per capita
                                          electricity use, kWh
------------------------------------------------------------------
Sorted by:  state
```

لإيجاد المتوسط والانحراف المعياري للاستخدام الفردي للكهرباء (*elcap*) قم بطباعة الأمر:

.summarize *elcap*

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       elcap |        51    13318.43    4139.328       6721      27457
```

الجدول أعلاه يوضح أيضاً عدد المشاهدات الموجودة للمتغير، ويعرض أعلى قيمة وأقل قيمة، وإذا قمنا بطباعة الأمر **summarize** بدون إضافة أي متغير، فسوف نحصل على المتوسطات والانحرافات المعيارية لكل متغير رقمي في البيانات.

لمشاهدة تفاصيل أكثر عن الملخصات الإحصائية، قم بطباعة الأمر

.summarize *elcap*, detail

Per capita electricity use, kWh

	Percentiles	Smallest		
1%	6721	6721		
5%	7434	7363		
10%	8286	7434	Obs	51
25%	10359	7467	Sum of Wgt.	51
50%	13388		Mean	13318.43
		Largest	Std. Dev.	4139.328
75%	16117	19477		
90%	17903	19896	Variance	1.71e+07
95%	19896	21590	Skewness	.7643711
99%	27457	27457	Kurtosis	4.161063

مخرجات الأمر summarize, detail: تحتوي على إحصائيات أساسية بالإضافة إلى المعلومات التالية:

المئينات: والملاحظ أن الربيع الأول (النسبة المئوية 25 = 10,359) والوسيط (النسبة المئوية 50 = 13,388) والربيع الثالث (النسبة المئوية 75 = 16,117)، وحيث إن العديد من العينات لايتم تقسيمها في شكل ربيعات أو أجزاء معيارية، فإن هذه النسب المئوية عبارة عن تقديرات.

وهناك في الجدول أربع أعلى قيم، وأربع أقل قيم، حيث يمكن أن تظهر القيم المتطرفة ضمن هذه القيم.

ومجموع الأوزان: الأمر **summarize** يسمح للأوزان التكرارية أو **fweight**، ولمزيد من الشرح قم بطباعة الأمر **help weight**.

التباين: مربع الانحراف المعياري (الاحتمال الأكثر هو أن الانحراف المعياري يساوي الجذر التربيعي للتباين).

الالتواء: وهو اتجاه ودرجة عدم التماثل في منحنى التوزيع الطبيعي، فالتوزيع الطبيعي المتماثل تماماً هو التوزيع الذي يكون فيه الالتواء يساوي صفرًا، أما الالتواء الموجب (ذيل المنحنى أطول في الجانب الأيمن) يعني أن الالتواء أكبر من الصفر، أما الالتواء السالب (ذيل المنحنى أطول في الجانب الأيسر) يعني أن الالتواء أصغر من الصفر.

التفرطح: وزن ذيل المنحنى، التوزيع الطبيعي (منحنى جاوس) يكون متماثلاً عندما يكون التفرطح يساوي 3، أما إذا كان منحنى التوزيع أطول من الطبيعي، (مدبب بشكل كبير) فإن التفرطح أكبر من 3، أما إذا كان التفرطح أقل من 3 فإن هذا يشير إلى ذيل أقل من الطبيعي.

الأمر tabstat: يعتبر بديلاً أكثر مرونة للأمر **summarize** حيث يمكننا تحديد الإحصائيات التي نريد حسابها، فمثلاً:

```
.tabstat elcap, stats(mean min max)
```

variable	mean	min	max
elcap	13318.43	6721	27457

باستخدام الأمر **tabstat** مع الخيار **by(*varname*)** يمكننا إنشاء جدول يحتوي على الملخصات الإحصائية لكل قيمة من قيم *varname*، المثال أدناه يقوم بإنشاء جدول للمتوسط، وأعلى وأقل قيمة لاستخدام للفرد الواحد للكهرباء بشكل منفصل لكل إقليم من الأقاليم الأمريكية الأربعة في التعداد السكاني، حيث إن استخدام الكهرباء يُعتبر منخفضاً في الشمال الشرقي Northeast ومرتفعاً في الجنوب South والنصف الغربي Midwest.

```
.tabstat elcap, stats(mean min max) by (region4)
```

Summary for variables: elcap
by categories of: region4 (Census Region (4))

region4	mean	min	max
Northeast	8746	7434	11759
Midwest	14151.5	10516	19477
South	16001.06	11343	21590
West	12206.92	6721	27457
Total	13318.43	6721	27457

بالإضافة إلى المتوسط **mean** وأقل قيمة **min** وأعلى قيمة **max**، هناك إحصائيات أخرى متوافرة مع الخيار **stats()** والأمر **tabstat** تتضمن المجموعة التي سبق استخدامها سابقاً مع الأمر **collapse** والأمر **graph bar**

(مثل count, sum, max, min, variance, sd, والمئينات من p1 إلى p99) وهناك خيارات إضافية أخرى يمكن من خلالها التحكم في شكل الجداول وتوصيفاتها. وللحصول على قائمة كاملة بهذه الخيارات، قم بطباعة الأمر **.help tabstat**

الإحصائيات التي تم إنشاؤها بواسطة الأمر **summarize**، أو الأمر **tabstat** تقوم بشرح وصفي للعينة، وخدمة لبعض الأغراض الأخرى، فإننا قد نقوم بتحديد فترة ثقة للاستدلال عن المجتمعات الكبيرة، وكما تم شرحه سابقاً، فإنه يمكننا الحصول على فترة ثقة 99% لمتوسط المتغير *elcap*

```
.ci elcap, level(99)
```

Variable	Obs	Mean	Std. Err.	[99% Conf.	Interval]
elcap	51	13318.43	579.6218	11766.32	14870.54

في بيانات العينة الموجودة لدينا، فإنه يمكننا التأكد بنسبة 99% أن متوسط المجتمع يكون في فترة الثقة إذا كانت قيمته بين 11,766 إلى 14,870 كيلووات في الساعة للفرد الواحد، وبشكل أكثر دقة فإنه في العديد من العينات العشوائية، فترات الثقة التي يتم إنشاؤها بهذه الطريقة يجب أن يكون متوسط مجتمعها نحو 95% عند اختيار العينة، فالخيار **level(99)** يحدد أن فترة الثقة تساوي 99%، وإذا قمنا بإهمال هذا الخيار، فإن الوضع الافتراضي للأمر **ci** هو اختيار 95% كفترة ثقة.

الخيارات الأخرى تسمح للأمر **ci** بحساب فترة ثقة محددة للمتغيرات التي تتبع توزيع ذي الحدين أو توزيع بواسون، والأمر المتعلق بهذه الحسابات هو **cii** الذي يقوم بحساب فترات الثقة مباشرة للتوزيع الطبيعي وتوزيع ذي الحدين وتوزيع بواسون وذلك من الملخصات الإحصائية. وللحصول على تفاصيل عن هذه الأوامر، قم بطباعة الأمر **help ci**.

تحليل البيانات الاستكشافي : Exploratory Data Analysis

الإحصائي Joh Tukey، جمع مجموعة من الأدوات للطرق الحديثة والقديمة لتحليل البيانات الاستكشافي. وهي تتضمن تحليل البيانات بطريقة استكشافية بدون

إجراء أي افتراضات غير ضرورية (انظـــر 1977 Tukey; Mosteller and Tukey 1983 ,1985) فرسومات الصندوق التي تم شرحها في الفصل (3) تعتبــر أكثــر الأشكال البيانية استعمالاً في تحليل البيانات الاستكشافي، وهناك طريقــة أخــرى وهي عرض الساق والورقة، وهي طريقة جغرافية لترتيب قيم البيانات، بحيث إن الأرقام الأولية تعتبر هي الساق، وباقي الأرقام لكل مشاهدة تعتبر الأوراق.

```
.stem elcap
  Stem-and-leaf plot for elcap (Per capita electricity use, kWh)

      6***  | 721
      7***  | 363,434,467,952
      8***  | 286,514,591,696,982,985
      9***  |
     10***  | 106,359,516,739
     11***  | 253,343,395,759
     12***  | 077,159,379,497,845,904
     13***  | 388,557,916,992
     14***  | 179,263,325,345,475,489,578
     15***  | 048,568
     16***  | 117,293,315,519,793
     17***  | 290,293,903
     18***  | 852
     19***  | 477,896
     20***  |
     21***  | 590
     22***  |
     23***  |
     24***  |
     25***  |
     26***  |
     27***  | 457
```

في هذا العرض، أقل قيمة لاستهلاك الكهرباء للفرد الواحد هي 6,721 (كاليفورنيا) حيث تظهر قيمة 721 كورقة للــساق ***6، وأعلــى قيمــة 27,457 (ويمينج) تظهر كورقة 457 للساق ***27، فالأمر **stem** يقــوم تلقائياً باختيار القيم للساق، ويمكنه تجاوز هذا باستخدام الخيار **line()**. ولمزيد من المعلومات حول هذا الخيار قم بطباعة الأمر **help stem**.

الأمر lv : يستخدم إحصائيات مرتبة لشرح التوزيع.

```
.lv elcap
```

#	51	Per capita electricity use, kWh				
					spread	pseudosigma
M	26		13388			
F	13.5	10437.5	13140	15842.5	5405	4131.039
E	7	8514	12903.5	17293	8779	3894.835
D	4	7467	13472	19477	12010	4098.322
C	2.5	7398.5	14070.75	20743	13344.5	3866.579
B	1.5	7042	15782.75	24523.5	17481.5	4369.45
	1	6721	17089	27457	20736	4689.655
					# below	# above
inner fence		2330		23950	0	1
outer fence		-5777.5		32057.5	0	0

في الجدول أعلاه، M يشير إلى الوسيط، F الربيعات (الربيعات باستخدام تقدير مختلف عن تقدير الربيعات الذي تم استخدامه في الأوامر **summarize, detail, tabsum**) أما E، D، C ... تشير إلى النقاط الفاصلة مثل 8/1، 16/1، 32/1 ... للتوزيع المتبقي خارج منطقة ذيل منحنى التوزيع، الأرقام بالعمود الثاني توضح المسافة أو العمق من أقرب نهاية لكل قيمة حرفية، أما الجزء الأوسط في الجدول فهو يتضمن ثلاثة أعمدة، العمود الأوسط يوضح متوسطات قمتين حرفيتين، إذا انتقلت هذه القيم بعيداً عن الوسيط كما حدث في المتغير *elcap* فإن هذا يعني أن التوزيع أصبح ذا التواء مرتفع، وأصبح يتجه أكثر نحو ذيل منحنى التوزيع، ويكون التشتت مختلفاً بين نوعين من القيم الحرفية. فمثلاً التشتت بين قيم F يساوي تقريباً المدى بين الربيعات؛ وأخيراً العمود الأخير في الجانب الأيمن pseudosigma يقوم بتقدير الانحراف المعياري الذي من المفترض إذا كانت القيم الحرفية تشرح مجتمع جاوس، قيم F في العمود الأخير pseudosigma في الجانب الأيمن أحياناً يطلق عليها الانحراف المعياري الوهمي (*PSD*) وهو يفحص عينة ما وقيمها المتطرفة لتقدير طبيعة التوزيعات المتماثلة:

1- مقارنة المتوسط مع الوسيط لفحص الالتواء بشكل كلي:

المتوسط > الوسيط	التواء موجب
المتوسط = الوسيط	متماثل
المتوسط < الوسيط	التواء سالب

2- إذا كان المتوسط والوسيط متشابهين، فهذا يشير إلى أن التوزيع متماثل. لذا فإن المقارنة بين الانحراف المعياري والانحراف المعياري الوهمي (*PSD*) تساعد في تقييم طبيعة ذيل منحنى التوزيع الطبيعي:

التوزيع الطبيعي PSD< ذيل منحنى التوزيع أكبر من الطبيعي

التوزيع الطبيعي $= PSD$ ذيل منحنى التوزيع طبيعي

التوزيع الطبيعي PSD> ذيل منحنى التوزيع أقل من الطبيعي

لنفرض أن F_1 وF_2 تشيران إلى الربع والثلاثة أرباع (تقريباً الربيع الأول والربيع الرابع) فإن المدى الربيعي IQR يساوي F_3 ناقصاً F_1، والانحراف المعياري الوهمي PSD يساوي IQR مقسوماً على 1.349.

الأمر **lv** يحدد أيضاً القيم المتطرفة تطرفاً حاداً وتطرفاً بسيطاً (هناك قيمة واحدة متطرفة تطرفاً بسيطاً في توزيع المتغير *elcap*) يمكننا أن نسمى القيمة x "قيمة متطرفة تطرفاً بسيطاً" عندما تبرز خارج الحد الداخلي وليس الحد الخارجي:

$$F_3 + 1.5IQR < x \le F_3 + 3IQR \text{ أو } F_1 - 3IQR \le x < F_1 - 1.5IQR$$

القيمة x"قيمة متطرفة تطرفاً حاداً" عندما تبرز خارج الحد الخارجي:

$$x > F_3 + 3IQR \quad \text{أو} \quad X < F_1 - 3IQR$$

الأمر **lv** يوضح هذه الحدود، كما يوضح أعداد القيم المتطرفة لكل نوع، القيم المتطرفة تطرفاً حاداً وهي القيم التي تقع خارج الحدود الخارجية وهي نادرة الحدوث (نحو 2 لكل مليون) في المجتمعات الطبيعية، تطبيقات محاكاة مونت كارلو تشير إلى أن وجود أي قيمة متطرفة في العينات التي حجمها n = 150 إلى n = 20,000 يفترض أن يكون دليلاً كافياً لرفض فرضية الاعتدال normality hypothesis عند $\alpha = 0.05$(Hamilton 1992b).

الأوامر **summarize, stem, lv** جميعها تؤكد أن توزيع العينة للمتغير *elcap* له التواء موجب، ولا يشبه منحنى التوزيع الطبيعي النظري. الجزء التالي من هذا الفصل، سوف يشرح اختبارات الاعتدال بشكل أكثر تفصيلاً، والتحويلات التي يمكن أن تقلل من التواء المتغيرات.

اختبارات الاعتدال والتحويلات :

Normality Tests and Transformations

العديد من الاختبارات الإحصائية تكون ذات كفاءة عندما يتم تطبيقها على متغيرات تتبع التوزيع الطبيعي. الجزء التالي سيشرح طرق استكشاف لفحص الاعتدال التقريبي، واستخدام أدوات الرسم البياني (المدرج التكراري، ورسم الصندوق، وشكل التماثل، وأشكال الربيعات) التي تم شرحها في الفصل 3، واختبارات الالتواء والتفرطح موضحاً لإحصائيات الالتواء والتفرطح عند استخدام الأوامر **summarize, detail** التي يمكنها تقييم فرضية العدم وهي أن العينة تم الحصول عليها من مجتمع موزع توزيعاً طبيعياً.

```
.sktest elcap
```

Skewness/Kurtosis tests for Normality

Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	joint adj chi2(2)	joint Prob>chi2
elcap	51	0.0223	0.0723	7.49	0.0236

الأمر sktest يرفض الاعتدال: المتغير *elcap* يظهر غير طبيعي بشكل ملحوظ حيث إن الالتواء (p=0.0223) بالرغم من أن التفرطح (p = 0.723) وكلتا الإحصائيتين (الالتواء والتفرطح) معاً (p = 0.0236).

اختبارات الاعتدال الأخرى تتضمن طرقاً لحساب إحصائية W لشابيرو ويلك Shapiro-Wilk W(**swilk**) وإحصائية W' لشابيرو فرانسيا Shapiro-Francia(**sfrancia**) (لمزيد من المعلومات قم بطباعة الأمر **help sktest**)، نموذج ستاتا لحساب اختبارات حلقات دورنيك هانسن Doornik-Hansen للاعتدال الأحادي والمتعدد متوافرة على الإنترنت (قم بطباعة الأمر **findit omninorm**).

التحويلات اللاخطيّة مثل الجذور التربيعية واللوغاريتمات يتم استخدامها عادةً لتغيير أشكال التوزيع، وذلك بهدف جعل التواء التوزيعات أكثر تماثلاً، وأقرب للتوزيع الطبيعي. التحويلات قد تساعد في جعل العلاقة بين

المتغيرات علاقة خطيّة (الفصلان 7 و8. الجدول (1.5) يعرض تعاقباً يسمى سلّم القوى (Tukey 1977) وهو يوضح دليلاً لاختيار التحويلات التي تقوم بتغيير شكل التوزيعات، المتغير *elcap* يعرض التواء موجباً بسيطاً لذا فإن جذره التربيعي قد يكون متماثلاً، ويمكننا إنشاء متغير جديد يساوي الجذر التربيعي للمتغير *elcap* وذلك بطباعة الأمر التالي:

```
.generate srelcap = elcap ^.5
```

وبدلاً من كتابة *elcap* ^.5 يمكننا كتابة **sqrt(*elcap*)**.

اللوغاريتمات هي عبارة عن تحويل آخر يمكنه تخفيض الالتواء الموجب، ولإنشاء متغير جديد(*logelcap*) يساوي اللوغاريتم الطبيعي للمتغير *elcap* يمكننا طباعة الأمر:

```
.generate logelcap = ln(elcap)
```

في سلم القوى وطرق التحويل ذات الصلة مثل بوكس-كوكس واللوغاريتمات فهي تأخذ مكان قوة 0، وهذه الطرق لها تأثير متوسط على شكل التوزيع، وهذا التأثير يتراوح ما بين 0.5 (الجذر التربيعي) و0.5- (عكس الجذر التربيعي) للتحويلات.

جدول (1.5): سلم القوى

التحويل	الصيغة	التأثير
المكعب	*new* = *old*^3	تخفيض الالتواء السالب الحاد
التربيع	*new* = *old*^2	تخفيض الالتواء السالب البسيط
بيانات خام	*old*	لا يوجد تغيير (بيانات خام)
الجذر التربيعي	*new* = *old*^.5	تخفيض الالتواء الموجب البسيط
اللوغاريتم أو (لوغاريتم 10)	*new* = ln(*old*) *new* = log10(*old*)	تخفيض الالتواء الموجب
سالب عكس الجذر التربيعي	*new* = - (*old*^-.5)	تخفيض الالتواء الموجب الحاد
المتبادلة السالبة	*new* = - (*old*^-1)	تخفيض الالتواء الموجب الحاد
مربع المتبادلة السالبة	*new* = - (*old*^-2)	تخفيض الالتواء الموجب الحاد
مكعب المتبادلة السالبة	*new* = - (*old*^-3)	تخفيض الالتواء الموجب الحاد

قمنا بأخذ القيم السالبة للنتيجة بعد رفع الأس لأقل من صفر. وللحفاظ على الترتيب الأصلي للبيانات، فإن أعلى قيمة في المتغير القديم *old* سوف يتم تحويلها إلى أعلى قيمة في المتغير الجديد *new* وهكذا، وعند احتواء المتغير *old* على قيمة سالبة أو صفر، فإنه من الضروري إضافة ثابت قبل إجراء عملية التحويل، فمثلاً إذا كان المتغير *arrests* يقوم بقياس عدد المرات التي تم فيها اعتقال شخص ما (وهناك قيمة صفر للعديد من الأشخاص) فإنه من الأفضل استخدام اللوغاريتم لإجراء التحويل.

```
.generate logarrest = ln(arrests + 1)
```

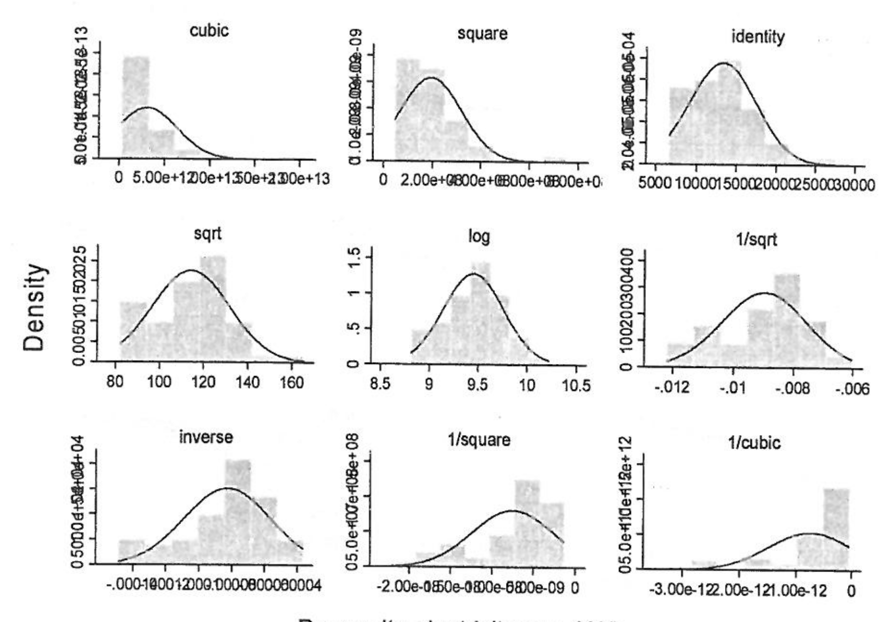
الأمر ladder : يجمع سلم القوى مع الأمر **sktest** للاعتدال، وهو يحاول استخدام كل قوة في السلم، ويفصح عما إذا كانت النتيجة غير طبيعية بشكل كبير. يمكن توضيح ذلك باستخدام الالتواء الموجب في المتغير *elcap*، والذي يحتوي على بيانات استهلاك الكهرباء لكل فرد بملف البيانات *electricity.dta*

```
.ladder elcap
```

Transformation	formula	chi2(2)	P(chi2)
cubic	elcap^3	44.12	0.000
square	elcap^2	26.24	0.000
identity	elcap	7.49	0.024
square root	sqrt(elcap)	1.21	0.547
log	log(elcap)	0.26	0.879
1/(square root)	1/sqrt(elcap)	2.36	0.307
inverse	1/elcap	4.87	0.088
1/square	1/(elcap^2)	10.67	0.005
1/cubic	1/(elcap^3)	17.51	0.000

الجذر التربيعي ومعكوس الجذر التربيعي ومعكوس التحويلات جميعها تقوم بتقريب التوزيعات التي لا تختلف بشكل كبير عن التوزيع الطبيعي. في هذا الصدد، فإن التحويلات هي عبارة عن تطويرات تتم إضافتها على البيانات الخام التي تختلف عن التوزيع الطبيعي بدرجة كبيرة (p=0.024). من الواضح أن اللوغاريتمات تعتبر أفضل خيار للتحويل الطبيعي، الشكل (1.5) والذي تم إنشاؤه بواسطة الأمر **gladder** يعطي دعماً مرئياً لهذه النتيجة، وذلك من خلال مقارنة المدرجات التكرارية لكل عملية تحويل إلى منحنى التوزيع الطبيعي.

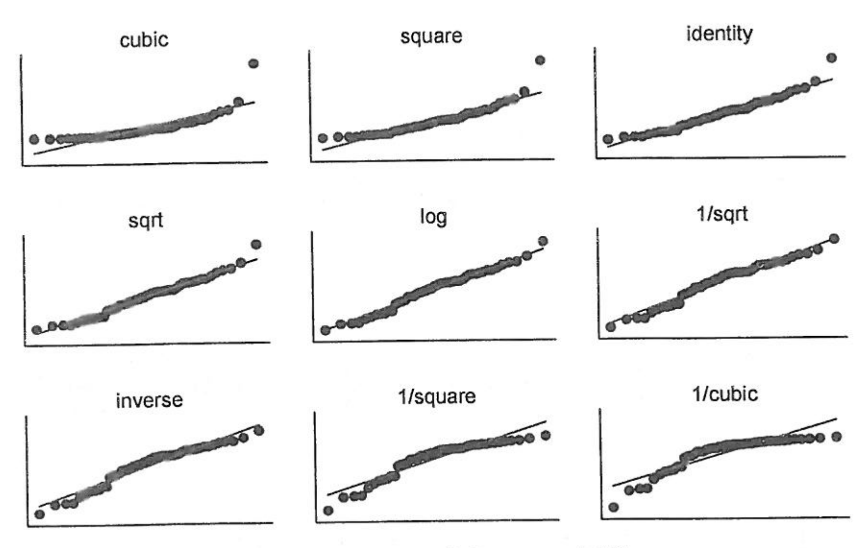
```
.gladder elcap
```



الشكل (1.5)

الشكل (2.5) يعرض مجموعة متقابلة من الرسومات البيانية لتحويلات سلّم القوى. تم إنشاء هذه الأشكال بواسطة أمر سلّم الربيعات **qladder**. (لمزيد من المعلومات عن **ladder, gladder, qladder** قم بطباعة الأمر **help ladder**)، ولجعل الرسومات البيانية الصغيرة أكثر وضوحاً في المثال أدناه، يمكننا جعل قياس التوصيفات والرموز في الرسم تظهر بحجم نسبته 25% من حجمها الأصلي، وذلك عن طريق استخدام الخيار **scale(1.25)** وتوصيفات المحاور (والتي قد يصعب قراءتها نظراً لتزاحمها) تم إخفاؤها عن طريق استخدام الخيار **ylabel(none) xlabel(none)**

```
.qladder elcap, scale(1.25) ylabel(none)
 xlabel(none)
```



الشكل (2.5)

هناك طريقة بديلة للتحويل يطلق عليها بوكس – كوكس Box-Cox، وهي توفر تدرجاً أفضل بين التحويلات، وآلية الاختيار فيما بين هذه المتغيرات (وهي سهلة للمحلل، ولكنها ليس دائماً الطريقة الأفضل)، الأمر **bcskew0** وجد بأن قيمة λ (لمدا) للتحويلات

أو:

$$y^{(\lambda)} = \ln(y) \qquad \lambda = 0$$

القيمة مثل $y^{(\lambda)}$ في العادة لها قيمة التواؤها يساوي صفرًا تقريباً، وبتطبيق ذلك على المتغير *elcap* يمكننا الحصول على متغير جديد يكون اسمه *belcap* وذلك عن طريق طباعة الأمر التالي:

```
.bcskew0 belcap = elcap, level(95)
```

Transform	L	[95% Conf.	Interval]	Skewness
(elcap^L-1)/L	.1451061	-.8268476	.8784035	3.75e-06

المعادلة $belcap=(elcap^{0.145}-1)/(0.145)$ تقــوم بالتحــويل لجعل التوزيع أكثر تماثــلاً ومقترباً من إحصـــائية الالتواء المطلوبة، بارامتر بــوكس – كوكس 0.145 = λ ليس من خيارات سلم القوى واللوغــاريتم (الــذي كــان مرفوعاً للأس 0)، وكانت فترة الثقة لــ λ تتضمن 0 (لوغاريتم) ولكــن لا تتضمن 1 (لا تغيير):

$$-0.827<\lambda<0.878$$

الفصل (8) يشرح طريقة كوكس – بوكس لنموذج الانحــدار بطريقــة أكثر تفصيلاً.

الجداول التكرارية.. والجداول التقاطعية الثنائية :

Frequency Tables and Two-Way Cross-Tabulations

الملخصات الإحصائية، والرسومات البيانية، والتحويلات التــي ســبق شرحها يمكن تطبيقها بشكل أساسي على المتغيرات القابلــة للقيــاس. أمــا المتغيرات الطبقية فتتطلب طرقاً أخرى في العادة تبدأ باستخدام جداول أحادية أو ثنائية. وبالعودة إلى بيانات مركز جرانيــت بــالملف *Granite2011_6.dta* هناك سؤال (*trackus*) عن ماذا كان الناس يعتقدون أن الولايات المتحدة تسير في الاتجاه الصحيح، أو أنها في الاتجاه الخطأ؟ بالرغم من أن هذا الــسؤال يبدو غامضاً، وتمت صياغته بطريقة غريبة، ولكنه سؤال تقليدي يُستخدم في استطلاعات الرأي بالولايات المتحدة لقياس المزاج العام للمواطنين، أغلبيــة المواطنين بولاية نيوهامبشير أظهروا تشاؤمهم عن وضع البلاد.

```
.tabulate trackus
```

US right direction or wrong track	Freq.	Percent	Cum.
Right direction	176	37.29	37.29
Wrong track	296	62.71	100.00
Total	472	100.00	

الأمر tabulate : يمكنه إنشاء جداول تكرارية للمتغيرات التي تحتوي على آلاف القيم. وقبل إنشاء جداول توزيع تكراري لمتغير معين يحتوي على العديد من القيم، فإننا نحتاج إلى تصنيف هذه القيم باستخدام الأمر **generate** مع الخيار **recode** والخيار **autocode** (لمزيد من المعلومات عن الأمر **generate** انظر الفصل 2 أو قم بطباعة الأمر **help generate**).

نقوم باستخدام الأمر **tabulate** يليه أسماء متغيرين اثنين لإنشاء جدول تقاطعي. فمثلاً لإنشاء جدول تقاطعي للمتغير *trackus* مع المتغير *educ* (مستوى التعليم لدى المشاركين بالدراسة) نقوم بطباعة الأمر:

```
.tabulate eductrackus
```

Highest degree completed	US right direction or wrong track Right dir	Wrong tra	Total
HS or less	36	71	107
Tech/some coll	34	76	110
College grad	49	93	142
Postgrad	56	52	108
Total	175	292	467

اسم المتغير الأول سوف يمثل الصفوف، والمتغير الثاني يمثل الأعمدة في الجدول، من الجدول أعلاه يمكننا أن نرى أن 71 من 107 مشارك في الدراسة مستواهم التعليمي الثانوية العامة أو أقل HS or less يعتقدون بأن الولايات المتحدة تسير في الاتجاه الخطأ.

ولكن السؤال: هل وجهات النظر *trackus* لها علاقة بالمستوى التعليمي؟ وللإجابة عن ذلك، يمكننا إجراء اختبار كاي تربيع χ^2، واختبار نسب الصف لأن المتغير *educ* – والذي يظهر في صفوف الجدول – يمثل المتغير المستقل في هذا الاختبار، الخيار **row** يحدد النسب المئوية للصفوف، والخيار **nof** يعني عدم إظهار التكرارات.

```
.tabulate educ trackus, row nof chi2
```

Highest degree completed	US right direction or wrong track Right dir	Wrong tra	Total
HS or less	33.64	66.36	100.00
Tech/some coll	30.91	69.09	100.00
College grad	34.51	65.49	100.00
Postgrad	51.85	48.15	100.00
Total	37.47	62.53	100.00

```
Pearson chi2(3) =  12.7549   Pr = 0.005
```

نحو 69% من المشاركين في الدراسة، لديهم مؤهل علمي تقني أو كلية Tech/some coll يعتقدون أن الولايات المتحدة في الاتجاه الخطأ، ولكن حملة الشهادات العليا postgrad يبدو أنهم أكثر تفاؤلاً، حيث إن نسبة 48% منهم يعتقدون نفس الاعتقاد. وبناءً على هذه العينة يمكننا أن نرفض فرضية العدم، وهي عدم وجود علاقة بين المتغير *educ* والمتغير *trackus* في مجتمع الدراسة بولاية هامبشير ($\chi^2 = 12.75, p = 0.005$).

الأمر **tabulate** به العديد من الخيارات امفيدة التي تساعد في إنشاء الجداول الثنائية. هذه الخيارات تتضمن اختبارات بديلة (اختبار الدقة لفيشر، معدل الإمكان χ^2)، ومقاييس العلاقات (قيمة جاما لجودمان وكروسكال γ، وكندال تاو أ τ_b ، ومعامل ارتباط كريمر V)، الخيار **missing** يحدد بأن القيم المفقودة يجب تضمينها في صفوف أو أعمدة الجدول، الأمر **tabulate** يمكنه حفظ التكرارات، وأسماء المتغيرات كمصفوفة، لمزيد من المعلومات عن هذه الخيارات قم بطباعة الأمر **help tabulate**.

أحياناً قد نحتاج إلى إعادة تحليل الجداول المنشورة بدون الرجوع إلى البيانات الأصلية. هناك أمر خاص وهو **tabi** (الجدول الفوري) يقوم بهذه المهمة. قم بطباعة تكرارات الخلايا في سطر الأمر مع فصل صفوف الجدول بعلامة " \ "، لشرح كيفية قيام الأمر **tabi** بإعادة إنشاء الجدول التقاطعي ما قبل السابق مباشرة من تكرارات الخلايا بدون الرجوع إلى أي بيانات:

```
.tabi 36 71 \ 34 76 \ 49 93 \ 56 52
```

row	col 1	2	Total
1	36	71	107
2	34	76	110
3	49	93	142
4	56	52	108
Total	175	292	467

```
Pearson chi2(3) =  12.7549   Pr = 0.005
```

وللقيام بنفس التحليل، وعرض النسب المئوية للصفوف، واختبار كاي تربيع χ^2

```
.tabi 36 71 \ 34 76 \ 49 93 \ 56 52, row nof chi2
```

row	col 1	2	Total
1	33.64	66.36	100.00
2	30.91	69.09	100.00
3	34.51	65.49	100.00
4	51.85	48.15	100.00
Total	37.47	62.53	100.00

```
Pearson chi2(3) =  12.7549   Pr = 0.005
```

الأمر tabi : يختلف عن الأمر **tabulate** في أنه لا يتطلب وجود أي بيانات في ذاكرة برنامج ستاتا، وعند إضافة الخيار **replace** يمكننا جعل الأمر **tabi** يقوم باستبدال أي بيانات في الذاكرة بالبيانات الجديدة التي تظهر في الجدول التقاطعي، الخيارات الإحصائية (**chi2, exact, nofreq**... الخ) تقوم بنفس المهام مع الأمر **tabi** التي قامت بها من قبل مع الأمر **tabulate**. ولمزيد من المعلومات قم بطباعة الأمر **help tabulate twoway**.

حتى الآن كل الأمثلة التي تم شرحها في هذا الجزء، لا تتضمن أوزاناً مرجحة، وكما تم شرحه سابقاً في الفصل (4) فإن الباحثين في الدراسات الاستقصائية في العادة يطبّقون الأوزان المرجحة بعناية فائقة، وذلك لجعل نتائج العينة ممثلة للمجتمع المستهدف، المتغير *censuswt* يمثل الأوزان المرجحة لبيانات استطلاع الرأي التي جمعها مركز جرانيت، وتم استخدام الأمر **svyset** للتأكيد بأن هذه الأوزان هي أوزان احتمالية.

```
.svyset [pw = censuswt]
```

الأوامر التي تبدأ بـــ : svy سوف تقوم بتطبيق الأوزان الاحتمالية svyset بشكل تلقائي، أما بالنسبة للأوامر الأخرى، فإنها تتجاهل الأوزان، وســوف نسرد بعض الأمثلة عن الأوزان الاحتمالية في الجداول الثنائية.

```
.svy: tab trackus
(running tabulate on estimation sample)

Number of strata   =        1          Number of obs       =       472
Number of PSUs     =      472          Population size     = 474.80568
                                       Design df           =       471

----------------------------
US right   |
direction  |
or wrong   |
track      |  proportions
-----------+----------------
Right di   |        .3696
Wrong tr   |        .6304
           |
Total      |            1
----------------------------
  Key:  proportions  =  cell proportions
```

```
.svy: tab eductrackus, row percent
(running tabulate on estimation sample)

Number of strata   =        1          Number of obs       =       467
Number of PSUs     =      467          Population size     = 469.25491
                                       Design df           =       466

-----------------------------------------------
Highest    | US right direction or wrong
degree     |           track
completed  | Right di  Wrong tr      Total
-----------+-----------------------------------
HS or le   |    34.33     65.67        100
Tech/som   |     24.5      75.5        100
 College   |    36.41     63.59        100
Postgrad   |    53.41     46.59        100
           |
   Total   |    37.26     62.74        100
-----------------------------------------------
  Key:  row percentages

  Pearson:
    Uncorrected   chi2(3)          =   21.3629
    Design-based  F(2.99, 1394.32)=    5.9918     P = 0.0005
```

في الجدول الذي يعرض الأوزان المرجحة، يمكننا أن نرى الفرق الكبير في التشاؤم بين المشاركين في الدراسة وبين الذين يحملون مؤهل المعهد التقني أو لم يُكمِلوا الجامعة Tech/some (75.5% يعتقدون أن الولايات المتحدة في الاتجاه الخطأ) والذين يحملون مؤهلات دراسات عليا (46.6% يعتقدون أن الولايات المتحدة في الاتجاه الخطأ)، التصميم بناءً على اختبار *F* يعطي نتائج مناظرة للجدول المرجح لاختبار كاي تربيع. اختبار *F* يؤكد أن العلاقة بين المتغير *educ* والمتغير *trackus* هي علاقة إحصائية ذات معنوية ($p = 0.005$).

الجداول المتعددة.. والجداول التقاطعية المتعددة :

Multiple Tables and Multi-Way Cross-Tabulations

عند العمل مع الدراسات الاستقصائية والبيانات الكبيرة، فإننا أحياناً نحتاج إلى التوزيعات التكرارية للعديد من المتغيرات المختلفة. وبدلاً من إنشاء كل جدول بشكل منفصل في كل مرة، يمكننا استخدام أمر آخر خاص وهو **tab1**.

```
.tab1 tparty Obama trackus
```

ولإنشاء جداول تكرارية أحادية لكل متغير من *tparty* وحتى المتغير *trackus* في هذه البيانات (في المرة الواحدة يمكنك استخدام 30 متغيراً كحد أقصى) قم بطباعة الأمر

```
.tab1 tparty-obama
```

وبالمثل، فإن الأمر **tab2** يقوم بإنشاء جداول ثنائية، فمثلاً الأمر التالي يقوم بإنشاء جداول تقاطعية ثنائية لكل متغير

```
.tab2 tparty obama trackus
```

والأمر **tab1** والأمر **tab2** يستخدمان الخيارات التي يستخدمها الخيار **tabulate**.

ولإنشاء جداول احتمالية متعددة، فإنه من الممكن استخدام الأمر **tabulate** مع وضع المحدد **by** قبل الأمر. فعلى سبيل المثال، لإنشاء جدول

تقاطعي أحادي عن المشاركين الذين قاموا بانتخاب الرئيس أوباما في سنة 2008 وعما إذا كان خريجو كليات أو لا، نقوم بطباعة الأمر التالي:

```
.tab obama college, col nof chi
```

Voted for Obama in 2008	College graduate Non colle	College	Total
No	61.44	41.45	50.68
Yes	38.56	58.55	49.32
Total	100.00	100.00	100.00

Pearson chi2(1) = 20.2966 Pr = 0.000

وهناك طريقة واحدة لإنشاء جدول تقاطعي ثلاثي للمتغير *Obama*، والمتغير *college* وعلاقتهما بمتغير الجنس *sex*، وذلك باستخدام الأمر **sort** والمُحدد **by**، حيث يقوم هذا المحدد بإنشاء جداول ثنائية بنفس تنسيق الجداول أعلاه، ولكن بشكل منفصل للذكور والإناث.

```
.sort sex
.by sex: tab obama college, col nof chi
```

-> sex = Male

Voted for Obama in 2008	College graduate Non colle	College	Total
No	69.81	44.88	56.22
Yes	30.19	55.12	43.78
Total	100.00	100.00	100.00

Pearson chi2(1) = 14.5888 Pr = 0.000

-> sex = Female

Voted for Obama in 2008	College graduate Non colle	College	Total
No	54.62	38.51	46.04
Yes	45.38	61.49	53.96
Total	100.00	100.00	100.00

Pearson chi2(1) = 7.2227 Pr = 0.007

العلاقة بين المتغيرين *Obama* و*college* ذات معنوية، وفي نفس الاتجاه للجدولين أعلاه، ولكن يظهر أن العلاقة أكثر قوة بين الرجال (حيث إن نسبة خريجي الجامعة تمثل 25 نقطة 30.19 إلى 55.12%) عن النساء (16 نقطة فرق، 45.38 إلى 61.79%).

هذه الطريقة يمكن استخدامها لإنشاء جداول أكثر تعقيداً، فمثلاً لإنشاء جدول تقاطعي رباعي للمتغير *obama* مع المتغير *college* مع جداول فرعية للرجال والنساء المتزوجين وغير المتزوجين، يمكننا طباعة الأمر التالي (لم يتم عرض نتائج هذا الأمر):

```
.sort sex married
.by sex married: tab Obama college, col nof chi
```

مثل هذا الجدول المتعدد يصنّف البيانات في عينات فرعية يكون فيها التباين أكثر قوة.

هناك طريقة أخرى لإنشاء الجداول المتعددة، فإذا كنا لا نحتاج إلى النسب المئوية أو الاختبارات الإحصائية، فإنه يمكننا استخدام الأمر العام لإنشاء الجداول وهو **table**. فهذا الأمر له عدة مزايا وعدة خيارات، تم عرض جزء بسيط منها فقط، ولإنشاء جدول ثنائي للمتغير *obama* مع المتغير *college* مع تكرارات في كل خلية نقوم بطباعة الأمر التالي:

```
.table obama college, contents(freq)
```

Voted for Obama in 2008	College graduate Non college	College
No	145	114
Yes	91	161

إذا قمنا بتحديد متغير طبقي ثالث، فسوف يتم إنشاء أعمدة فرعية في جدول ثلاثي كما يلي:

```
.table obama college sex, contents(freq)
```

Voted for Obama in 2008	Gender and College graduate — Male —		— Female —	
	Non college	College	Non college	College
No	74	57	71	57
Yes	32	70	59	91

الجداول الأكثر تعقيداً تتطلب استخدام الخيار **by()** والذي يمكنه استخدام أربعة متغيرات فرعية أخرى، لذا فإن الأمر **table** يمكنه إنشاء جدول لسبعة متغيرات (عمود واحد وصف واحد وعمود فرعي واحد وأربعة صفوف فرعية) ويتم إنشاء ذلك الجدول كما يلي:

```
.table obama college sex,contents(freq) by(married)
```

Respondent married and Voted for Obama in 2008	Gender and College graduate — Male —		— Female —	
	Non college	College	Non college	College
No				
No	34	12	35	22
Yes	15	22	39	38
Yes				
No	40	45	36	35
Yes	17	48	20	53

الأمثلة أعلاه استخدمت الأمر **table** وقامت بوضع التكرارات في خلايا الجداول، ولكن الأمر **table** يتيح لنا إنشاء ملخصات إحصائية، فمثلاً الجدول الرباعي للمتغيرات *obama* × *college* × *sex* × *married* يحتوي في كل خلية على متوسط العمر *age* لمجموعة من الخصائص، حيث نرى أن 34 لم يتخرجوا في الكلية non-college رجال غير متزوجين لم يصوّتوا لأوباما ومتوسط أعمارهم 46.6 سنة.

```
.table obama college sex, contents(mean age)
 by(married)
```

Respondent married and Voted for Obama in 2008	Gender and College graduate Male Non college	Male College	Female Non college	Female College
No				
No	46.63636	46.91667	60.64706	60
Yes	55.6	53.45454	63.21053	61.78378
Yes				
No	56.075	55.92857	58.2	52.35484
Yes	59	55.87234	53.21053	53.80769

الخيار **contents()** مع الأمر **table** يحدد الإحصائيات التي تحتويها خلايا الجدول. الخيارات لا تتضمن التكرارات أو المتوسطات فقط، وإنما تتضمن أيضاً الانحراف المعياري، وأعلى قيمة، وأقل قيمة، والوسيط، والمدى، والنسب المئوية، وملخصات أخرى. وللحصول على قائمة كاملة بهذه الخيارات قم بطباعة الأمر **table help**. الجزء التالي من هذا الفصل، سوف يشرح بعض الاحتمالات الخاصة بالملخصات الإحصائية بالجداول.

جداول المتوسطات والوسيط والملخصات الإحصائية الأخرى :

Tables of Means, Medians and Other Summary Statistics

الأمر **tabulate** يقوم بإنشاء جداول للمتوسطات، والانحراف المعياري على شكل فئات للمتغيرات. بالنسبة للأمثلة المتبقية في هذا الفصل، سوف نعود لاستخدام بيانات استهلاك الكهرباء بالولايات المتحدة، الأمر **tabulate** يعرض طريقة واحدة لمشاهدة الملخصات الإحصائية لمعدل استهلاك الكهرباء لكل فرد (*elcap*) لكل إقليم في التعداد السكاني بالولايات المتحدة (*region9*).

```
.tabulate region9, summ(elcap)
```

Census Division (9)	Summary of Per capita electricity use, kWh Mean	Std. Dev.	Freq.
New Engla	8417.1667	532.95419	6
Mid Atlan	9403.6667	2176.4139	3
E N Centr	12726.2	2274.5595	5
W N Centr	15169.571	2172.0833	7
S Atlanti	15011.889	2810.0798	9
E S Centr	17948.25	2475.1953	4
W S Centr	16279.5	1965.8288	4
Mountain	13877.5	5723.3327	8
Pacific	9534	3073.2846	5
Total	13318.431	4139.3277	51

كما يمكننا استخدام الأمر **tabulate** لإنشاء جداول ثنائية للمتوسطات كما في هذا المثال، باستخدام تقسيمات الأقاليم الموجودة بالتعداد السكاني للولايات المتحدة:

```
.tabulate region9 region4, summ(elcap) mean
```

Means of Per capita electricity use, kWh

Census Division (9)	Census Region (4) Northeast	Midwest	South	West	Total
New Engla	8417.1667	.	.	.	8417.1667
Mid Atlan	9403.6667	.	.	.	9403.6667
E N Centr	.	12726.2	.	.	12726.2
W N Centr	.	15169.571	.	.	15169.571
S Atlanti	.	.	15011.889	.	15011.889
E S Centr	.	.	17948.25	.	17948.25
W S Centr	.	.	16279.5	.	16279.5
Mountain	.	.	.	13877.5	13877.5
Pacific	.	.	.	9534	9534
Total	8746	14151.5	16001.059	12206.923	13318.431

الخيـــار mean : في الأمر أعلاه يحدد بأن الجدول يجب أن يحتوي على المتوسطات فقط وإذا لم نقم باستخدام هذا الخيار فإننا سوف نحـــصل علـــى جدول ضخم يحتوي على المتوسطات والانحراف المعياري والتكرارات في كل خلية.

الأمر table : مرن ويستخدم عند إنشاء جداول لسبعة متغيرات، ويحتوي الجدول على المتوسطات والانحراف المعيـــاري، والمجـــاميع، والوســـيط، وإحصائيات أخرى. ولشرح ذلك، فإن الجدول أدناه عبارة عن جدرل أحادي يعرض المتوسط، والانحراف المعياري لاستهلاك الكهرباء للفرد الواحد، كما يعرض أيضاً المدى الربيعي للمجتمع لكل إقليم بالتعداد السكاني.

```
.table region9, contents
 (mean elcap sd elcap median pop iqr pop)
```

Census Division (9)	mean(elcap)	sd(elcap)	med(pop)	iqr(pop)
New England	8417.17	532.9542	1322	2521
Mid Atlantic	9403.67	2176.414	12702	10586
E N Central	12726.2	2274.56	9884	5053
W N Central	15169.6	2172.083	2853	4490
S Atlantic	15011.9	2810.08	5774	7682
E S Central	17948.3	2475.195	4559.5	1910
W S Central	16279.5	1965.829	4142	11506
Mountain	13877.5	5723.333	2380	2618
Pacific	9534	3073.285	3831	5365

معدل استهلاك الكهرباء للفرد الواحد يتباين نتيجة لعامل من اثنين، فهو معدل منخفض يساوي 8,417 كيلووات/ساعة في نيوإنجلاند إلى معدل مرتفع 17,948 كيلووات/ساعة في وسط الجنوب الغربي (هذا يتضمن الولايات المنتجة للنفط وهي تكساس ولويزيانا وأوكلاهوما). ومن ناحية أخرى، فإن أعلى تباين حدث في الولايات الجبلية. حيث إن الانحراف المعياري (5,723 كيلووات/ساعة) أعلى بعشر مرات عنه في ولاية نيوإنجلاند (533 كيلووات/ساعة).

الخيار contents(): في الأمر **table**، يحدد الإحصائيات التي يجب أن تظهر في كل خلية لكل متغير. الإحصائيات التي يمكن إدخالها تتضمن أعلى قيمة وأقل قيمة والمجموع والنسبة والمئينات وعدة أنواع من الأخطاء المعيارية. للحصول على قائمة بهذه الخيارات قم بطباعة الأمر **help table**.

استخدام الأوزان التكرارية: Using Frequency Weights

الأوامر **summarize, tabulate, tabe** وعدة أوامر أخرى، يمكن استخدامها مع الأوزان التكرارية التي تشير إلى عدد المشاهدات المتكررة. فمثلاً المتوسط والإحصائيات الأخرى لاستهلاك الفرد للكهرباء بالولايات المتحدة كما يلي:

```
.summ elcap
```

Variable	Obs	Mean	Std. Dev.	Min	Max
elcap	51	13318.43	4139.328	6721	27457

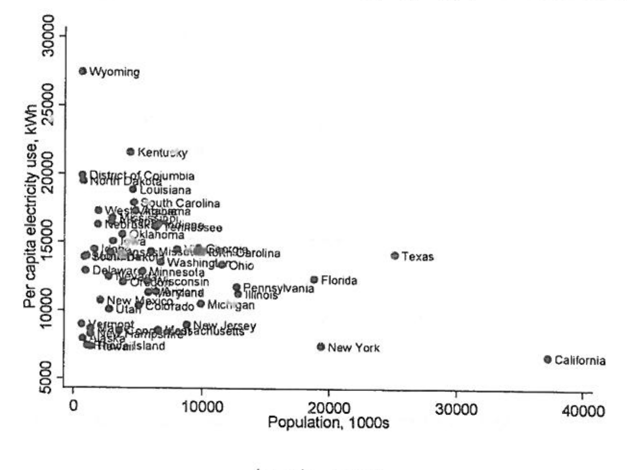
المتوسط 13,318 كيلووات/ساعة يمثل متوسط استهلاك الكهرباء في 51 ولاية (بما فيها ضاحية كولومبيا بالعاصمة واشنطن). ويتم اعتبار كل ولاية وحدة واحدة، ولاية وايومنج Wyoming التي بها أقل عدد سكان (564 ألف نسمة) وأعلى معدل لاستهلاك الكهرباء للفرد الواحد (27,457 كيلووات/ساعة). أما في ولاية كاليفورنيا California التي بها أكبر عدد سكان (37 مليوناً) وأقل معدل لاستهلاك الكهرباء للفرد الواحد (6,721 كيلووات/ساعة) كل ولاية لها نفس الوزن عند حساب المتوسط للولايات 51، ولتوضيح المتوسط لكل فرد في الولايات المتحدة ككل، فإنه يجب القيام بوزن عدد السكان.

```
.summ elcap [fweight = pop]
```

Variable	Obs	Mean	Std. Dev.	Min	Max
elcap	308746	12112.69	3519.441	6721	27457

المتوسط المرجح لاستهلاك الكهرباء لعدد السكان بالولايات المتحدة (12,114 كيلووات/ساعة) أقل من متوسط 51 ولاية (13,318 كيلووات/ساعة) بسبب كثرة عدد السكان الذين يعيشون في الولايات ذات الاستهلاك المنخفض للكهرباء مثل كاليفورنيا California ونيويورك New York عن عدد السكان في الولايات ذات الاستهلاك المرتفع للكهرباء مثل وايومنج Wyoming وولاية كنتاكي Kentucky (الشكل 3.5).

```
.graph twoway scatter elcap pop, mlabel(state)
```



الشكل (3.5)

المتوسط الموزون بعدد السكان لاستهلاك الكهرباء للفرد الواحد يمكن شرحه كمتوسط لعدد سكان الولايات المتحدة بالكامل والبالغ 309 ملايين نسمة، ويجب ملاحظة أنه لا يمكننا القيام بحساب انحراف معياري موزون أو أعلى قيمة أو أقل قيمة، وذلك بسبب أن أغلب الإحصائيات الفردية لايمكن حسابها من بيانات موزونة، لأنها مجموعة أصلاً. لذلك علينا التعامل بحذر مع الأوزان، لأنها ذات معنى مع نوع معين من التحليل، ولكنها نادراً ما يكون لها معنى مع البيانات ككل عند استخدام أنواع مختلفة من التحليل.

الأوزان التكرارية تقوم بنفس العمل مع الأمر **tabulate** والأمر **table**. الأمر أدناه يقوم بحساب المتوسط الموزون لعدد السكان لكل إقليم، وبهذا يمكننا أن نأخذ في الاعتبار الولايات ذات الكثافة السكانية الكبيرة، حيث يمكننا أن نرى من الجدول أدناه أن أصغر متوسط لاستهلاك الكهرباء في إقليم الباسفيك Pacific.

```
.table region9 [fweight = pop], contents(mean elcap) row
```

Census Division (9)	mean(elcap)
New England	8486.42
Mid Atlantic	9127.38
E N Central	12444.7
W N Central	14374.9
S Atlantic	13891.4
E S Central	17819.2
W S Central	15092.9
Mountain	11816.9
Pacific	8089.04
Total	12112.7

الخيار row : يحدد أن الصف الأخير يلخص الجدول ككل، فالمتوسط العام في هذا الجدول يساوي (12,112.7 كيلووات/ساعة) هو نفس المتوسط الذي حصلنا عليه سابقاً عند استخدام الأمر **summarize**.

الفصل السادس

تحليل التباين وطرق المقارنة الأخرى

Anova and Other Comparison Methods

تحليل التباين ANOVA، يتضمن مجموعة من الطرق لاختبار الفرضيات حول الاختلافات بين المتوسطات، ويمكن تطبيق هذا التحليل على نطاق واسع يمتد من التحليلات البسيطة التي يمكننا من خلالها المقارنة بين متوسط المتغير y من خلال فئات المتغير x ، وحتى التحليلات المعقدة مع المتغيرات التصنيفية المتعددة والمتواصلة للمتغيرات. اختبار t للفرضيات المتعلقة بمتوسط فردي (عينة واحدة) أو بزوج من المتوسطات (عينتان) ترتبط بالأشكال الابتدائية لتحليل ANOVA.

الاختبارات اللامعلمية التي تعتمد على الرتب، مثل اختبار مان وتني، واختبار كروسكال واليز، لها طريقتها المختلفة لمقارنة التوزيعات. هذا الاختبار يقوم بافتراضات ضعيفة حول القياس وشكل التوزيع والانتشار. ولكن تظل هذه الاختبار صالحة إذا توافر لها عدد كبير من الشروط أكثر من تلك التي يتطلبها تحليل التباين وعناصره المعلمية. أحياناً قد يقوم المحللون باستخدام الاختبارات المعلمية واللامعلمية لفحص ما إذا كانت النتائج تسير في اتجاه متشابه، وإذا اختلفت نتائج الاختبارات المعلمية عن نظيراتها اللامعلمية، فيجب محاولة الكشف عن هذا الاختلاف، ومحاولة معرفة أسبابه.

الأمر anova: هو أحد أوامر ستاتا النموذجية، فهو مثله مثل الأوامر الأخرى، يعتبر مرناً بدرجة كبيرة، ويتضمن عدداً كبيرًا من النماذج. فالأمر **anova** يتوافق مع تحليل التباين الأحادي والمتعدد، كما أنه متوافق مع التغاير (ANCOVA) للتصميمات المتوازنة وغير المتوازنة بما فيها الخلايات

المفقودة، كما أنه متوافق مع التصميمات العامليّة، والتصميمات التجريبية المتشابكة، والتصميمات المختلطة، وتصاميم القياسات المتكررة. أحد الأوامر التابعة الأخرى هو الأمر **predict** الذي يقوم بحساب القيم المتوقعة لعدة أنواع من البواقي، وعدة أخطاء معيارية، والإحصاءات التشخيصية. ويتم استخدام هذا الأمر بعد الأمر **anova**. الأمر الآخر التابع هو **test** ويقوم بحساب الاختبارات التي يحددها المستخدم لاختبار فرضية العدم. الأمر **test** والأمر **predict** يعملان بنفس الطريقة مع أوامر ستاتا الأخرى، التي تتوافق مع النماذج مثل الأمر **regress** الذي (سيتم شرحه في الفصل 7).

خيارات قوائم ستاتا أدناه تؤدي إلى القيام بأغلب الإحصائيات التي تم شرحها في هذا الفصل وهي كما يلي:

Statistics > Summaries, tables, & tests > Classical tests of hypotheses
Statistics > Summaries, tables, & tests > Nonparametric tests of hypotheses
Statistics > Linear models and related > ANOVA / MANOVA
Statistics > Postestimation > Predictions residuals, etc.
Graphics > Twoway graph (scatter, line etc.)

أمثلة عن الأوامر : Example Commands

```
.anova y x1 x2
```

يقوم بحساب تحليل التباين الثنائي، موضحاً الاختلافات بين متوسطات المتغير *y* من خلال تصنيفات المتغير *x1* والمتغير *x2*.

```
.anova y x1 x2x1#x2
```

يقوم بحساب تحليل التباين العاملي ذي الاتجاهين، الذي يتضمن التأثيرات الأساسية والتفاعلية (*x1#x2*) للمتغيرات التصنيفية *x1* و*x2*، كما يمكن تحديد نفس النموذج بالضبط باستخدام الرمز العاملي عن طريق الأمر **anova *y x1##x2*** حيث إن (##) لا يسمح فقط بتفاعل *x1##x2* وإنما يسمح أيضاً بأي نوع من أنواع تأثيرات المستوى المنخفض والتأثيرات الأساسية، بما فيها هذه المتغيرات. (في هذا المثال البسيط، هناك تأثيرات أساسية لكل من *x1* و *x2*).

```
.anova yx1##x2##x3
```

يقوم بحساب تحليل التباين العاملي ذي الثلاثة اتجاهات، وهذا يتضمن ثلاثة اتجاهات لتفاعل المتغيرات *x1##x2##x3* بالإضافة إلى كل التفاعلات الثنائية (*x1#x2, x1#x3, x2#x3*) والتأثيرات الأساسية (*x1, x2, x3*).

```
.anova reading curriculm / teacher|curriculm /
```

يقوم بصياغة نموذج تجريبي متشابك مناسب لاختبار تأثيرات ثلاثة أنواع من المناهج *curriculm* على قدرة الطلبة على القراءة (*reading*)، والمتغير *teacher* المعلّم متشابك مع المنهج *curriculm*(*teacher|curriculum*) لأن مجموعات مختلفة من المعلّمين تم تخصيصهم لكل منهج، دليل المستخدم *Base Reference Manual* يوضح مع الأمثلة النماذج التجريبية المتشابكة الأخرى للتباين، كما يتضمن أيضاً تصميم القطع المنفصل.

```
.anova headache subject medication,
 repeated(medication)
```

يقوم بجعل نموذج التباين للقياسات المتكررة متناسباً لاختبار تأثيرات ثلاثة أنواع من علاج الصداع (*medication*) على شدّة الصداع (*headache*)، العينة تحتوي على 20 شخصاً يعانون من صداع متكرر. وكل شخص استخدم أنواع العلاج الثلاثة في أوقات مختلفة أثناء الدراسة.

```
.anova y x1 x2 c.x3 c.x4x2#c.x3
.regress
```

يقوم بحساب التغاير (ANCOVA) لأربعة متغيرات مستقلة، منها اثنان تصنيفيان (*x1, x2*) واثنتان متصلتان (*x3, x4*) وبما في ذلك التفاعل بين *x2#c.x3*، الأمر الذي يليه وهو **regress** بدون إضافة أي متغيرات وهو يقوم بإنشاء جدول لنتائج الانحدار.

```
.kwallis y, by(x)
```

يقوم بحساب اختبار كروسكال ويلز لاختبار فرضية العدم للمتغير *y*، الذي له توزيع مرتب متشابه لعدد الفئات *k* للمتغير *x* حيث إن ($k>2$).

```
.oneway y x
```

يقوم بحساب تحليل التباين الأحادي (ANOVA) مختبراً الفروقات بين المتوسطات للمتغير *y* مع فئات المتغير *x*، نفس التحليل – مع اختلاف جدول النتائج – يمكن القيام به باستخدام الأمر **anova** *y x*.

```
.oneway y x, tabulate scheffe
```

يقوم هذا الأمر، بحساب تحليل التباين ANOVA متضمناً جدول مخرجات لمتوسطات العينة، واختبارات المقارنة المتعددة لشافيه Scheffé.

```
.ranksum y, by(x)
```

يقوم هذا الأمر، بحساب اختبار مجموع الرتب لويلكوكسن (كما يُعرف أيضاً باختبار *U* مان وتني) لفرضية العدم والتي يكون فيها المتغير *y* له توزيعات رتب متشابهة لفئات المتغير الثنائي *x*. وإذا فرضنا أن توزيعات الرتب لها نفس الشكل فهذا يُضيف اختبارًا عما إذا كان الوسيطان للمتغير *y* متساويين.

```
.serrbar ymean se x, scale(2)
```

يقوم برسم أعمدة بيانية تمثل الخطأ المعياري من بيانات المتوسطات، المتغير *ymean* يتضمن مجموعة المتوسطات للمتغير *y*، والمتغير *se* الخطأ المعياري، المتغير *x* يمثل قيم فئات المتغير *x*، والخيار **scale(2)** يحدد بأن امتداد الأعمدة البيانية يجب أن يكون 2± للخطأ المعياري حول كل متوسط (الوضع الافتراضي هو 1±).

```
.signrank y1 = y2
```

يقوم بإجراء اختبار ويلكوسن Wilcoxon للأزواج المتطابقة ورتب الإشارة الذي يختبر تساوي توزيعات الرتب للمتغير *y1* والمتغير *y2*، ويمكننا اختبار ما إذا كان الوسيط للمتغير *y1* يختلف عن قيمة ثابتة مثل 23.4 وذلك بطباعة الأمر **signrank *y1*=23.4**

```
.signtest y1 = y2
```

يختبر تساوي الوسيط للمتغير *y1* والمتغير *y2* (بافتراض أن بيانات المتغيرين متشابهة من حيث القياس ونفس مشاهدات العينة)، استخدام الأمر **signtest *y1* = 5** سوف يعرض اختبار الإشارة لفرضية العدم، وهي أن وسيط المتغير *y1* يساوي 5.

```
.ttest y = 5
```

يقوم بإجراء اختبار t لعينة واحدة لفرضية العدم التي تفترض بأن متوسط المجتمع للمتغير y تساوي 5.

```
.ttest y1 = y2
```

يقوم بإجراء اختبار t لعينة واحدة (الاختلاف المترابط) لفرضية العدم التي تفترض بأن متوسط المجتمع للمتغير $y1$ يساوي نظيره للمتغير $y2$، الوضع الافتراضي لهذا الأمر يفترض بأن البيانات مترابطة، فعند استخدام بيانات غير مترابطة (تم الحصول على بيانات المتغيرين $y1$ و$y2$ من عينتين مستقلتين) فيجب إضافة الخيار **unpaired**.

```
.ttest y, by(x) unequal
```

يقوم هذا الأمر بإجراء اختبار t لعينتين لاختبار فرضية العدم التي تفترض بأن متوسط المجتمع للمتغير y هو نفسه لتصنيفات المتغير x، لا يجب أن يُفترض بأن المجتمعات لها تباين متساو (بدون استخدام الخيار **unequal** فإن الوضع الافتراضي للأمر **ttest** هو اعتبار التباين متساوياً للمجتمعات).

اختبارات العينة الواحدة : One-Sample Tests

يبدو أن اختبار t للعينة الواحدة له عدة تطبيقات منها:

1- اختبار ما إذا كان متوسط العينة $\bar{y}$ يختلف بشكل ملحوظ عن القيمة المفترضة μ_0.

2- اختبار ما إذا كانت متوسطات المتغير y_1 والمتغير y_2 وهما متغيران تم قياسهما بنفس المقياس عند جمع قيم المشاهدات ولكنهما يختلفان عن بعضهما، وهذا يكافئ اختبار ما إذا كان متوسط متغير نتيجة الاختلاف تم إنشاؤه بواسطة طرح y_1 من y_2 يساوي صفر.

يمكننا استخدام نفس المعادلات للتطبيقات أعلاه بالرغم من أن التطبيق الثاني يبدأ بمعلومات عن متغيرين اثنين بدلاً من متغير واحد.

ملف البيانات *writing.dta*: يحتوي على معلومات تم جمعها لتقييم دورات الكتابة بالجامعة بناءً على معالجة الكلمات (دراسة Nash and Schwartz 1987)، القياسات – مثل عدد العبارات الكاملة في الوقت المحدد – تم جمعها قبل وبعد إتمام الطلبة للدورة. الباحثون يريدون معرفة ما إذا كانت القياسات بعد الدورة شهدت أي تطور.

```
.use C:\data\writing.dta, clear
.describe
```

```
Contains data from C:\data\writing.dta
  obs:            24                          Nash and Schwartz (1987)
 vars:             9                          2 Jul 2012 06:11
 size:           216
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
id              byte   %8.0g       slbl       Student ID
preS            byte   %8.0g                  # of sentences (pre-test)
preP            byte   %8.0g                  # of paragraphs (pre-test)
preC            byte   %8.0g                  Coherence scale 0-2 (pre-test)
preE            byte   %8.0g                  Evidence scale 0-6 (pre-test)
postS           byte   %8.0g                  # of sentences (post-test)
postP           byte   %8.0g                  # of paragraphs (post-test)
postC           byte   %8.0g                  Coherence scale 0-2 (post-test)
postE           byte   %8.0g                  Evidence scale 0-6 (post-test)
-------------------------------------------------------------------------------
Sorted by:
```

بافتراض أننا نعرف أن الطلبة خلال السنة الماضية كان معدل إكمال الطالب للعبارة هو 10 عبارات، وقبل اختبار مدى تطور مهاراتهم خلال الدورة قد نحتاج إلى معرفة مهاراتهم عند بداية الدورة. بعبارة أخرى نحن نحتاج إلى معرفة ما إذا كان متوسط العينة قبل الدورة (*preS*) يختلف بدرجة كبيرة عن متوسط الطلبة السابقين (10). وللحصول على اختبار t لعينة واحدة لفرضية العدم $H_0: \mu = 10$ نقوم بطباعة الأمر التالي:

```
.ttest preS = 10
```

```
One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    preS |      24    10.79167    .9402034    4.606037    8.846708    12.73663
------------------------------------------------------------------------------
    mean = mean(preS)                                             t =   0.8420
Ho: mean = 10                                    degrees of freedom =       23

    Ha: mean < 10               Ha: mean != 10                 Ha: mean > 10
 Pr(T < t) = 0.7958         Pr(|T| > |t|) = 0.4084          Pr(T > t) = 0.2042
```

الرمز **Pr(T<t)**: يعني "احتمال أن قيمة توزيع t أقل من قيمة t المشاهدة إذا كانت H_0 صحيحة" وهذا في حالة اختبار الاحتمال لطرف واحد، أما اختبار الاحتمال لطرفين لقيمة t المطلقة، فإن ذلك يعني Pr(|T|>|t|) = 0.4084 وحيث إن الاحتمال مرتفع فليس لدينا أي سبب لرفض $H_0: \mu = 10$ ويجب ملاحظة أن الأمر **ttest** يقوم بالاختبار معتبراً أن فترة الثقة للمتوسط تساوي 95%. وفترة الثقة هذه تتضمن قيمة فرضية العدم 10، ويمكننا أن نرى فترة ثقة مختلفة – 90% مثلاً – إذا استخدمنا الخيار **level(90)** مع الأمر أعلاه.

النظير اللامعلمي – اختبار التصميم – يستخدم توزيعاً ثنائياً لاختبار فرضيات عن قيم وسيط أحادية. فمثلاً يمكننا اختبار ما إذا كان الوسيط للمتغير *preS* يساوي 10، الأمر **signtest** يوضح بأنه لا يوجد سبب لرفض فرضية العدم.

```
.signtest preS = 10

        Sign test

                sign |    observed    expected
        -------------+------------------------
            positive |          12          11
            negative |          10          11
                zero |           2           2
        -------------+------------------------
                 all |          24          24

        One-sided tests:
          Ho: median of preS - 10 = 0 vs.
          Ha: median of preS - 10 > 0
              Pr(#positive >= 12) =
                 Binomial(n = 22, x >= 12, p = 0.5) =  0.4159

          Ho: median of preS - 10 = 0 vs.
          Ha: median of preS - 10 < 0
              Pr(#negative >= 10) =
                 Binomial(n = 22, x >= 10, p = 0.5) =  0.7383

        Two-sided test:
          Ho: median of preS - 10 = 0 vs.
          Ha: median of preS - 10 != 0
              Pr(#positive >= 12 or #negative >= 12) =
                 min(1, 2*Binomial(n = 22, x >= 12, p = 0.5)) =  0.8318
```

الأوامر مثل الأمر **ttest** والأمر **signtest** تتضمن الذيل الأيمن والذيل الأيسر لمنحنى التوزيع والاحتمالات الثنائية، على خلاف توزيعات *t* المتماثلة التي استخدمها الأمر **ttest** فإن التوزيعات الثنائية التي يستخدمها الأمر **signtest** لها شكل مختلف للذيل الأيسر والذيل الأيمن لمنحنى التوزيع. في هذا المثال الاحتمال الثنائي هو المهم لدينا، وذلك لأننا نريد اختبار ما إذا كانت بيانات الطلبة بالملف *writing.dta* تختلف عن فرضية العدم وهي أن الوسيط يساوي 10.

الآن سوف نقوم باختبار التطور خلال الدورة وذلك باختبار فرضية العدم التي تفترض بأن متوسط عدد العبارات التي يمكن إكمالها قبل وبعد الدورة (متوسط المتغير *preS* ومتوسط المتغير *postS* متساويان)، الأمر **ttest** يقوم بهذا الاختبار، ويوضح بأن هناك تطوراً ملحوظاً.

```
.ttest postS = preS
```

```
Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
   postS |      24      26.375    1.693779    8.297787    22.87115    29.87885
    preS |      24    10.79167    .9402034    4.606037    8.846708    12.73663
---------+--------------------------------------------------------------------
    diff |      24    15.58333    1.383019    6.775382    12.72234    18.44433
------------------------------------------------------------------------------
     mean(diff) = mean(postS - preS)                              t =  11.2676
 Ho: mean(diff) = 0                              degrees of freedom =       23

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 1.0000       Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

وحيث إننا نتوقع "تطوّرًا" وليس "اختلافاً" فقط في متوسطات المتغيرين *preS* و*postS*، فإن الاختبار الأحادي هو الاختبارالمناسب، احتمال الذيل الأيمن المعروض يقترب من الصفر. وهذا يعني أن متوسط إكمال الطلبة للعبارات تطوّر بدرجة كبيرة. وبناءً على بيانات المثال، فإننا على ثقة بدرجة 95% بأن مهارات الطلبة في كتابة العبارات الكاملة قد زادت بمعدل ما بين 12.7 و18.4 عبارة.

اختبارات t الاعتيادية: تفترض بأن المتغيرات تتوزع توزيعاً طبيعياً حول متوسطاتها. هذه الافتراضية في العادة ليست ذات أهمية بالغة، لأن هذه الاختبارات تُعتبر متوسطة الثقة، ولكن إذا كان عدم الاعتدال يتضمن قيماً متطرفة حادة – وهذا يحدث في العينات الصغيرة – فإنه من الأفضل الانتقال إلى الوسيط بدلاً من المتوسطات، واستخدام اختبار لامعلمي لا يفترض الاعتدال. فعلى سبيل المثال، اختبار ويلكوكسن لرتب الإشارة Wilcoxon signed-rank test يفترض أن التوزيع متماثل ومستمر فقط، وتطبيق اختبار الرتب على بيانات المثال السابق، سوف يؤدي إلى الحصول على نفس النتيجة التي وجدها الأمر **ttest**، وهي أن هناك تطورًا ملحوظاً للطلبة في إكمال العبارات. وحيث إن الاختبارين وجدا نفس النتيجة، فإنه بالإمكان إقرار ذلك بثقة أكبر.

```
.signrank postS = preS

Wilcoxon signed-rank test

        sign |      obs   sum ranks    expected
-------------+---------------------------------
    positive |       24         300         150
    negative |        0           0         150
        zero |        0           0           0
-------------+---------------------------------
         all |       24         300         300

unadjusted variance     1225.00
adjustment for ties       -1.63
adjustment for zeros       0.00
                     ----------
adjusted variance       1223.38

Ho: postS = preS
             z =   4.289
    Prob > |z| =   0.0000
```

اختبارات العينتين : Two-Sample Tests

بقية هذا الفصل، سوف تشرح أمثلة من بيانات دراسة استقصائية تم جمعها من طلبة الجامعة، وقام بالدراسة Ward و Ault في سنة 1990.

```
.use "C:\data\student2.dta", clear
.describe
```

```
Contains data from C:\data\student2.dta
  obs:           243                          Student survey (Ward 1990)
 vars:            18                          2 Jul 2012 06:11
 size:         5,346
-----------------------------------------------------------------------------
              storage  display    value
variable name   type   format     label      variable label
-----------------------------------------------------------------------------
id              int    %8.0g                 Student ID
year            byte   %9.0g      year       Year in college
age             byte   %8.0g                 Age at last birthday
gender          byte   %9.0g      s          Gender (male)
relig           byte   %8.0g      v4         Religious preference
drink           byte   %9.0g                 33-point drinking scale
gpa             float  %9.0g                 Grade Point Average
grades          byte   %8.0g      grades     Guessed grades this semester
greek           byte   %9.0g      greek      Belong to fraternity or sorority
live            byte   %8.0g      v10        Where do you live?
miles           byte   %8.0g                 How many miles from campus?
study           byte   %8.0g                 Avg. hours/week studying
athlete         byte   %9.0g      athlete    Are you a varsity athlete?
employed        byte   %8.0g      employ     Are you employed?
allnight        byte   %8.0g      allnight   How often study all night?
ditch           byte   %8.0g      times      How many class/month ditched?
hsdrink         byte   %9.0g                 High school drinking scale
aggress         byte   %9.0g                 Aggressive behavior scale
-----------------------------------------------------------------------------
Sorted by:  year
```

نحو 19% من الطلبة ينتمون إلى جمعية الطلبة الذكور أو جمعية الطلبة الإناث بالجامعة. في الحرم الجامعي هذه الجمعيات وأعضاؤها يتم الإشارة إليهم بأنهم "يونانيون" ليس بسبب جنسيتهم ولكن لأن أغلب أسماء هذه الجمعيات تتألف من حروف يونانية.

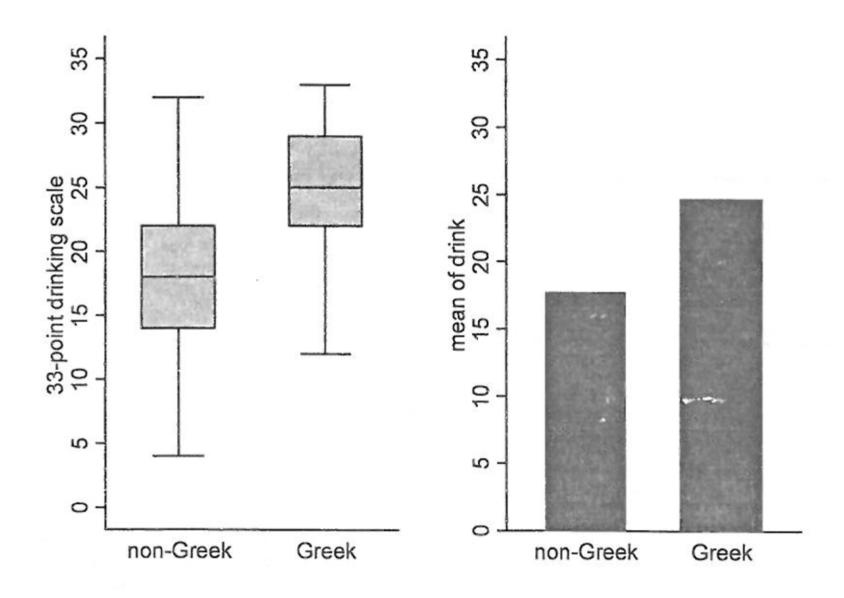
```
.tabulate greek

  Belong to |
 fraternity |
or sorority |      Freq.     Percent        Cum.
------------+-----------------------------------
  non-Greek |        196       80.66       80.66
      Greek |         47       19.34      100.00
------------+-----------------------------------
      Total |        243      100.00
```

وهناك متغير آخر وهو *drink* يقيس كم مرة وإلى أي مدى الطلبة يتناولون الكحول، والقياس عبارة عن 33 نقطة، الشائعات في الحرم الجامعي قد تقود إلى الظن بأن أعضاء الجمعيات الذكور والإناث يميلون للاختلاف عن بقية الطلبة في سلوكهم عند تناول الكحول، رسم الصندوق يُقارِن بين

الوسيط لقيم المتغير *drink* للأعضاء وغير الأعضاء، ورسم أعمدة بيانية يُقارن بين المتوسطات. الشكلان يبدو أنهما متفقان على قبول مثل هذه الشائعات، الشكل (1.6) يجمع بين الشكلين وبعد استخدام الخيار **ylabel(0(5)25)** لجعل قياسات المحور العمودي متناسبة للشكلين معاً.

```
.graph box drink, over(greek) ylabel(0(5)35)
 saving(fig06_01a)
.graph bar (mean) drink, over(greek)
 ylabel(0(5)35) saving(fig06_01b)
.graph combine fig06_01a.gph fig06_01b.gph,
 col(2) iscale(1.05)
```



الشكل (1.6)

الأمر **ttest** الذي تم تطبيقه سابقاً على عينة واحدة واختبارات الاختلاف المترابطة يمكن تطبيقها على عينتين. في هذا التطبيق الشكل العام لتركيبة الأمر هو **ttest *measurement*, by(*categorical*)** فمثلاً:

```
.ttest drink, by(greek)
```

```
Two-sample t test with equal variances

   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
non-Gree |     196     17.7602    .4575013    6.405018    16.85792    18.66249
   Greek |      47     24.7234    .7124518    4.884323    23.28931     26.1575
---------+--------------------------------------------------------------------
combined |     243      19.107     .431224    6.722117    18.25756    19.95643
---------+--------------------------------------------------------------------
    diff |             -6.9632    .9978608               -8.928842   -4.997558
------------------------------------------------------------------------------
    diff = mean(non-Gree) - mean(Greek)                           t =  -6.9781
Ho: diff = 0                                     degrees of freedom =      241

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

يمكن أن نلاحظ بأن اختبار t يقوم على افتراض تساوي التباين. ولكــن التباين في عينة الأعضاء الذكور والإناث – في المثال أعلاه – يبدو أنه أقل بطريقة ما، حيث إنهم يتشابهون بدرجة كبيرة في سلوكهم مع الطلبــة غيــر الأعضاء في الجمعية، ولإجراء اختبار مشابه بــدون افتــراض أن التبــاين متساو نقوم بإضافة الخيار **unequal** كما يلي:

```
.ttest drink, by(greek) unequal
```

```
Two-sample t test with unequal variances

   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
non-Gree |     196     17.7602    .4575013    6.405018    16.85792    18.66249
   Greek |      47     24.7234    .7124518    4.884323    23.28931     26.1575
---------+--------------------------------------------------------------------
combined |     243      19.107     .431224    6.722117    18.25756    19.95643
---------+--------------------------------------------------------------------
    diff |             -6.9632     .8466965               -8.645773   -5.280627
------------------------------------------------------------------------------
    diff = mean(non-Gree) - mean(Greek)                           t =  -8.2240
Ho: diff = 0                     Satterthwaite's degrees of freedom =    88.22

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

التصحيح لتساوي التباين لا يُغيّر النتيجة الأساســية وهــي أن Greeks (الأعضاء الذكور والإناث) وnon-Greek (غير الأعضاء) يختلفــان بــشكل

ملحوظ، يمكننا فحص هذه النتيجة من خلال استخدام اختبار مـــان وتنـــي U اللامعلمي الذي يُعرف كذلك باسم اختبار مجموع الرتب لويلكوكـــسن الـــذي يفترض بأن ترتيب التوزيعات له نفس شكل المنحنى، اختبار مجموع الرتب يشير إلى أننا نستطيع أن نرفض فرضية العدم، وهي أن قيم الوســـيط فـــي المجتمع متساوية.

```
.ranksum drink, by(greek)
```

```
Two-sample Wilcoxon rank-sum (Mann-Whitney) test

       greek |      obs    rank sum    expected
-------------+---------------------------------
   non-Greek |      196       21111       23912
       Greek |       47        8535        5734
-------------+---------------------------------
    combined |      243       29646       29646

unadjusted variance   187310.67
adjustment for ties     -472.30
                      ----------
adjusted variance     186838.36

Ho: drink(greek==non-Greek) = drink(greek==Greek)
             z =  -6.480
    Prob > |z| =   0.0000
```

تحليل التباين الأحادي (ذي الاتجاه الواحد) (ANOVA) :

One-Way Analysis of Variance (ANOVA)

تحليل التباين (ANOVA) يعتبر طريقة أخرى عامة أكثر من اختبارات t لاختبار الاختلافات بين المتوسطات، أبسط حالات التباين – وهي التحليـــل الأحادي للتباين – يختبر ما إذا كانت المتوسطات للمتغير y تختلف بين فئات المتغير x. التباين الأحادي يمكن حسابه عن طريق استخدام الأمـــر **oneway** مع الصيغة العامة **oneway** ***measurementcategorical*** فمثلاً:

```
.oneway drinkgreek, tabulate
```

```
 Belong to  |
 fraternity |  Summary of 33-point drinking scale
or sorority |        Mean   Std. Dev.       Freq.
------------+------------------------------------
  non-Greek |   17.760204   6.4050179         196
      Greek |   24.723404   4.8843233          47
------------+------------------------------------
      Total |   19.106996   6.7221166         243

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      1838.08426      1   1838.08426     48.69     0.0000
 Within groups      9097.13385    241   37.7474433
------------------------------------------------------------------------
    Total           10935.2181    242   45.1868517

Bartlett's test for equal variances:  chi2(1) =   4.8378  Prob>chi2 = 0.028
```

الخيار tabulate، يقوم بإنشاء جدول للمتوسطات والانحرافات المعيارية، بالإضافة إلى جدول تحليل التباين نفسه، تحليل التباين الأحادي مع متغير ثنائي x يكافئ اختبار t لعينتين، وتكون قيمة إحصائية F لهذا الاختبار تساوي مربع إحصائية t، الأمر **oneway** يوفر خيارات أكثر، ويعالج البيانات بسرعة أكثر، ولكن ينقصه الخيار **unequal** الذي يساعد فرضية تساوي التباين.

الأمر oneway: يقوم باختبار فرضية تساوي التباين باستخدام اختبار بارتليت Bartlett's χ^2 فعندما تكون احتمالية بارتليت منخفضة، فهذا يعني أن فرضية تساوي التباين غير صحيحة. وفي هذه الحالة، يجب عدم الوثوق في نتائج اختبار F للتباين. في المثال أعلاه الأمر **oneway *drink belong*** احتمالية بارتليت $p = 0.028$، مشيراً إلى وجود شكوك حول صلاحية تحليل التباين ANOVA.

القيمة الحقيقية لتحليل التباين الأحادي لا تكمن في قدرته على المقارنة بين عينتين، بل في قدرته على المقارنة بين ثلاثة متوسطات أو أكثر. فمثلاً، يمكننا اختبار ما إذا كان متوسط سلوك الطلبة يتفاوت لكل سنة في الكلية. في الجدول أدناه كلمة "freshman" تشير إلى طلبة السنة الأولى وليس بالضرورة أن يكونوا من الذكور.

```
.oneway drinkyear, tabulate scheffe
```

Year in college	Summary of 33-point drinking scale Mean	Std. Dev.	Freq.
Freshman	18.975	6.9226033	40
Sophomore	21.169231	6.5444853	65
Junior	19.453333	6.2866081	75
Senior	16.650794	6.6409257	63
Total	19.106996	6.7221166	243

Analysis of Variance

Source	SS	df	MS	F	Prob > F
Between groups	666.200518	3	222.066839	5.17	0.0018
Within groups	10269.0176	239	42.9666008		
Total	10935.2181	242	45.1868517		

Bartlett's test for equal variances: chi2(3) = 0.5103 Prob>chi2 = 0.917

Comparison of 33-point drinking scale by Year in college (Scheffe)

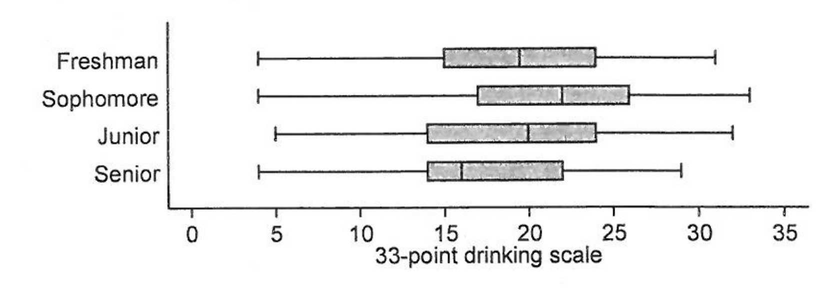
Row Mean-Col Mean	Freshman	Sophomor	Junior
Sophomor	2.19423 0.429		
Junior	.478333 0.987	-1.7159 0.498	
Senior	-2.32421 0.382	-4.51844 0.002	-2.80254 0.103

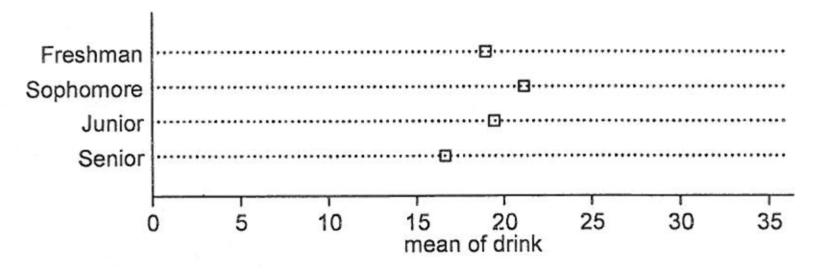
يمكننا رفض الفرضية القائلة بأن المتوسطات متساوية ($p = 0.0018$) ولكن لا يمكننا رفض فرضية تساي التباين ($p = 0.917$)، النتيجة الثانية تُعتبر أخبارًا جيدة حول صلاحية ANOVA.

رسم الصندوق الأفقي (**graph hbox**) في الشكل (2.6) يدعم هذه النتيجة، حيث يعرض تبايناً متشابهاً لكل فئة. وفي الشكل البياني يظهر رسم الصندوق مُدمجاً مع رسم بياني لشكل الانتشار (**graph dot(mean)**) يوضح المتوسطات لكل فئة. الرسم البياني الموحّد للشكلين يوضح بأن الفروقات بين قيم الوسيط (في أعلى الشكل) والفروقات بين المتوسطات (في أسفل الشكل) كلاهما تغيرا بطريقة متشابهة. الرسم البياني لشكل الانتشار يوضح نفس النقاط التي يوضحها رسم الأعمدة البيانية، فكلاهما يُستخدمان للمقارنة

البصرية بين الملخصات الإحصائية لمتغير واحد أو عدة متغيرات، تركيبة الأوامر والخيارات لرسم الأعمدة البيانية والرسم النقطي متشابهان فكلاهما يتضمن خيارات الملخصات الإحصائية. لمزيد من التفاصيل قم بطباعة الأمر **help graph dot**.

```
.graph hbox drink, over(year) ylabel(0(5)35)
 saving(fig06_02a)
.graph dot (mean) drink, over(year)
 ylabel(0(5)35, grid)
 marker(1, msymbol(Sh)) saving(fig06_02b)
.graph combine fig06_02a.gph fig06_02b.gph,
 row(2) iscale(1.05)
```





الشكل (2.6)

الخيار scheffe: (اختبارات المقارنة المتعددة لشافيه Scheffé) مع الأمر **oneway** يقوم بإنشاء جدول يعرض الفروقات بين كل زوج من المتوسطات، فمتوسط المتغير *freshman* يساوي 18.975 ومتوسط المتغير *sophomore* يساوي 21.6923 لذا فسوف يكون عبارة عن *freshman−sophomore* وهو

يساوي 21.6923 − 18.975 = 2.19423 وهـو ليس بعيـداً عن الصفر ($p = 0.429$)، وبالنسبة لهذه المقارنات في الجدول أعلاه الفـرق الوحيـد ذو المعنوية هو الفرق بين المتغيـر *senior* والمتغيـر *sophomore* 16.6508 − 21.1692 = 4.5184− ($p = 0.002$). لذلك فإن النتيجـة النهائيـة هـي أن المتوسطات الأربعة ليست نفسها التي تظهر من المقارنـات بـين الطلبـة بالسنوات المتقدمة بالجامعة seniors (الذين يشربون بنوع من الاتزان)، وبين طلبة السنة الثانية sophomores (الذين يُفرِطُونَ في الشرب).

الأمر oneway، يمكنه استخـدام خيارات متعـددة للمقارنـة وهـي **scheffe, bonferroni, sidak**. (لمعرفة تعريفات هذه الخيارات انظـر دليـل المستخدم*Base Reference Manual*) اختبار Scheffé يظل صالحاً فـي حالـة وجود عدد كبير من الشروط بالرغم من أن هذا الاختبار أقل حساسية أحياناً.

اختبار كروسكال والس (kwallis): وتعميم عينة *K* على مجموع رتـب عينة ثنائية يعتبران من الاختبارات اللامعلمية، وهما بديل لتحليل التباين لعينة واحدة. فاختبار كروسكال والس يختبر فرضية العدم القائلـة بتـساوي قـيم الوسيط للمجتمع.

```
.kwallis drink, by(year)

Kruskal-Wallis equality-of-populations rank test
```

year	Obs	Rank Sum
Freshman	40	4914.00
Sophomore	65	9341.50
Junior	75	9300.50
Senior	63	6090.00

```
chi-squared =     14.453 with 3 d.f.
probability =      0.0023

chi-squared with ties =    14.490 with 3 d.f.
probability =      0.0023
```

النتائج أعلاه ($p = 0.0023$) تتفق مع نتائج **oneway** بأن هنـاك فروقـات معنوية في المتغير *drink* لكل سنة في الكلية. وبصفة عامة، فـإن اختبـار

كروسال والس يعتبر أكثر أماناً من تحليل التباين ANOVA في حالة وجود سبب للشك في فرضيات تحليل التباين التي تفترض بتساوي التباين أو الاعتدال أو في حالة وجود مشاكل بسبب القيم المتطرفة، الأمر **kwallis** يشبه الأمر **ranksum** حيث إنهما يعتمدان على فرضية ضعيفة، وهي تماثل شكل الرتب في التوزيعات لكل مجموعة، نظرياً فإن الأمرين **ranksum** و**kwallis** يُفترض أن يقوما باستخراج نتائج متشابهة عند تطبيقهما على عينتين متشابهتين، ولكن في الواقع فإن هذا يكون صحيحاً فقط إذا كانت البيانات لا تحتوي على أي روابط. الأمر **ranksum** يحتوي على الطريقة الدقيقة للتعامل مع الروابط، وهي طريقة مفضّلة لمشاكل العينتين.

تحليل التباين ذي الاتجاهين والمتعدد :

Two- and *N*-Way Analysis of Variance

تحليل التباين ذي الاتجاه الواحد يختبر كيف أن متوسطات المتغير *y* تختلف خلال فئات متغير واحد آخر وهو *x*. تحليل التباين المتعدد يقوم بهذا التحليل للتعامل مع فئتين أو أكثر من فئات المتغير *x*. فعلى سبيل المثال، قد نحتاج إلى إعادة النظر كيف أن سلوك الطلبة عند تناولهم للكحول يختلف ليس فقط بين الطلبة والطالبات أعضاء الجمعيات، ولكن أيضاً الاختلاف في الجنس. وسوف نبدأ باختبار المتوسطات في جدول ثنائي كما يلي:

```
.table greekgender, contents(mean drink) row col
```

Belong to fraternity or sorority	Gender (male) Female	Male	Total
non-Greek	16.51724	19.5625	17.7602
Greek	22.44444	26.13793	24.7234
Total	17.31343	21.31193	19.107

نتائج هذه العينة، توضح بأن الذكور أكثر تناولاً للكحول من الإناث، وأعضاء جمعيات الطلبة من الذكور والإناث أكثر تناولاً للكحول من غير

الأعضاء في هذه الجمعيات. الفرق بين Greek/non-Greek يبدو متشابهاً بين الذكور والإناث.

الأمر **anova**، والذي يمكن استخدامه لتحليل التباين المتعدد، يمكنه اختبار الفروقات المعنوية للأعضاء والجنس (سوف تتم كتابته على هذا الشكل *greek#gender*).

```
.anova drink greek gender greek#gender
```

```
                   Number of obs =       243     R-squared     =  0.2221
                   Root MSE      =   5.96592     Adj R-squared =  0.2123
```

Source	Partial SS	df	MS	F	Prob > F
Model	2428.67237	3	809.557456	22.75	0.0000
greek	1406.2366	1	1406.2366	39.51	0.0000
gender	408.520097	1	408.520097	11.48	0.0008
greek#gender	3.78016612	1	3.78016612	0.11	0.7448
Residual	8506.54574	239	35.5922416		
Total	10935.2181	242	45.1868517		

في هذا المثال، نتائج تحليل التباين ذي الاتجاهين توضح بأن هناك تأثيرات أساسية ذات معنوية لأعضاء الجمعية ($greek(p = 0.0000$ والجنس ($gender(p = 0.0008$ ولكن تفاعلهما يساهم بدرجة قليلة في النموذج ($p = 0.7448$) لأن هذا التفاعل لا يمكن تمييزه عن الصفر، وقد نشير إلى أنه يُناسب نموذج أبسط بدون مصطلح التفاعل.

لإضافة أي مصطلح تفاعل مع الأمر **anova** نحدد أسماء المتغيرات وربطها مع بعضها بالرمز # (أو ## للتفاعل العاملي). إذا لم يكن عدد المشاهدات لكل مجموعة في قيم المتغير x هي نفسها، فإن (هذا الشرط يُطلق عليه البيانات المتوازنة) فقد يكون من الصعب تفسير التأثيرات الرئيسة في نموذج ما يتضمن تفاعلات. هذا لا يعني أن التأثيرات الرئيسة في مثل هذه النماذج ليست مهمة. تحليل الانحدار قد يساعد في تفسير نتائج تحليل التباين المعقدة كما سوف نرى لاحقاً في الجزء التالي.

المتغيرات العاملية وتحليل التغاير (ANCOVA) :

Factor Variables and Analysis of Covariance (ANCOVA)

الأمر **anova**، والعديد من أوامر التقدير ببرنامج ستاتا، تسمح بتحديد المتغيرات المستقلة وكتابتها في متغير عامليّ. فقبل كتابة اسم متغير مستقل نضع قبله الرمز .**i** والذي يحدد لبرنامج ستاتا أن هذا المتغير يتضمن متغيرًا تنبؤياً (ثنائياً) لمستويات متغير تصنيفي، لأن كل تصنيف له تفرعاته الثنائية، المتغيرات التصنيفية والتي يسبقها الرمز .**i** يجب أن تحتوي على أعداد صحيحة موجبة من 0 إلى 32,740 والأمر **anova** سوف يقوم بشكل افتراضي باعتبار كل المتغيرات المستقلة متغيرات تصنيفية. لذا قم بطباعة الأمر التالي:

```
.anova drink greek year greek#year
```

كما يمكن تنفيذ نفس النموذج بطباعة الأمر:

```
.anova drink i.greek i.year i.greek#i.year
```

```
                 Number of obs =      243     R-squared     =  0.2265
                 Root MSE      = 5.99962      Adj R-squared =  0.2034
```

Source	Partial SS	df	MS	F	Prob > F
Model	2476.29537	7	353.756482	9.83	0.0000
greek	1457.93596	1	1457.93596	40.50	0.0000
year	217.492051	3	72.4973502	2.01	0.1127
greek#year	148.508479	3	49.5028264	1.38	0.2510
Residual	8458.92273	235	35.9954159		
Total	10935.2181	242	45.1868517		

وكما هي العادة مع تحليل التباين ANOVA يمكننا الحصول على نظرة مباشرة للنموذج الضمني، وذلك بإعادة شرح التحليل على أنه تحليل انحدار، وبرنامج ستاتا يقوم بذلك بسهولة، فقط قم بطباعة الأمر **regress** مباشرة بعد أمر **anova** بدون إضافة أي متغيرات.

```
.regress
```

Source	SS	df	MS
Model	2476.29537	7	353.756482
Residual	8458.92273	235	35.9954159
Total	10935.2181	242	45.1868517

Number of obs	=	243
F(7, 235)	=	9.83
Prob > F	=	0.0000
R-squared	=	0.2265
Adj R-squared	=	0.2034
Root MSE	=	5.9996

drink	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
1.greek	7.805556	3.162076	2.47	0.014	1.575917	14.03519
year						
2	1.138889	1.322791	0.86	0.390	-1.467156	3.744934
3	.3648776	1.268844	0.29	0.774	-2.134884	2.864639
4	-3.043501	1.295774	-2.35	0.020	-5.596319	-.4906827
greek#year						
1 2	-.7859477	3.586922	-0.22	0.827	-7.852579	6.280683
1 3	-3.614878	3.58588	-1.01	0.314	-10.67945	3.4497
1 4	1.643501	3.778548	0.43	0.664	-5.800655	9.087657
_cons	18.19444	.9999363	18.20	0.000	16.22446	20.16443

لاحظ بأن مجاميع مربعات اختبار F وR^2 وتفاصيل أخـــرى متطابقــة لمكافئ تحليلات الأمر **anova** والأمر **regress**، كما أن جدول **regress** يوفر تفاصيل أكثر من أمر **anova**، ويمكننا مشاهدة كل قيمة من قيم المتغير *year* حيث تم معاملة هذا المتغير كمؤشر للتنبؤ، حيث إننا نرى بأن طلبة الـــسنة الأولى من فئة مختلطة في هذا الجدول.

لذا فإن المُعَامِلات في السنة الثانية والثالثة والرابعة تظهر عكس مُعَامِل السنة الأولى، ففي السنة الثانية الطلبة غير الأعضاء non-Greek كان معـــدل تناولهم المشروبات الكحولية أكثر من (+1.14)، في حين أن هذا المعدل كان منخفضاً جداً في السنة الرابعة (3.04−) مقارنـــة بطلبـــة الـــسنة الأولـــى، مُعاملات متغير السنة *year* متعلقة بمُعامِلات لمتغيرات وهمية تم ترميزهـــا بالرقم 1 لسنة معينة و0 لأي سنة أخرى، معاملات المتغيـــر *greek* متعلقـــة بمعامل متغير وهمي تم ترميزها بإعطاء رقم 1 للأعضاء Greek و0 لغيـــر الأعضاء non-Greek.

تحليل التغاير (ANCOVA)، يمتد لعدد *N* طريقة لتحليل التباين ANOVA ليشمل خليطاً من متغيرات *x* التصنيفية والمتصلة، المحدد .c والذي يسبق عددًا من الأوامر، يقوم بتحديد متغير مستقل معين كمتغير مستمر، وتتم معاملة قيمه كقياسات بدلاً من معاملتها كقيم مستقلة، وتكون تحت فئات معينة، قد يمكننا معاملة متغير *year* كمتغير متصل.

```
.anova drink i.greek c.year i.greek#c.year

                    Number of obs =      243     R-squared     =  0.1965
                    Root MSE      = 6.06334      Adj R-squared =  0.1864

             Source |  Partial SS    df       MS           F     Prob > F
         -----------+----------------------------------------------------
              Model |  2148.60352     3  716.201174      19.48     0.0000
                    |
              greek |  186.474269     1  186.474269       5.07     0.0252
               year |  147.628787     1  147.628787       4.02     0.0462
         greek#year |  .203073456     1  .203073456       0.01     0.9408
                    |
           Residual |  8786.61458   239  36.7640778
         -----------+----------------------------------------------------
              Total |  10935.2181   242  45.1868517
```

```
.regress

      Source |       SS       df       MS              Number of obs =     243
-------------+------------------------------           F(  3,   239) =   19.48
       Model |  2148.60352     3  716.201174           Prob > F      =  0.0000
    Residual |  8786.61458   239  36.7640778           R-squared     =  0.1965
-------------+------------------------------           Adj R-squared =  0.1864
       Total |  10935.2181   242  45.1868517           Root MSE      =  6.0633

------------------------------------------------------------------------------
       drink |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1.greek |   6.776657    3.00897     2.25   0.025     .8491681    12.70415
        year |  -1.103421   .4068558    -2.71   0.007    -1.904902   -.3019392
             |
greek#c.year |
           1 |   .0789217   1.061895     0.07   0.941    -2.012947     2.17079
             |
       _cons |   20.69328   1.164985    17.76   0.000     18.39833    22.98823
------------------------------------------------------------------------------
```

الشكل الجديد للنتائج بعد معاملة السنة *year* كمتغير متصل (c.*year*) بدلاً من متغير تصنيفي (i.*year*) يجعل النموذج أكثر بساطة مع درجات حرية

أعلى، ولكن R^2 المعدلة توضح بأن هذه النسخة من النتائج غير متناسبة (0.1864 مقابل 0.2034)، الإصدار التصنيفي للنتائج يوضح بأن معدل تناول الشراب مرتفع في السنة الثانية (1.14+) مقارنة مع السنة الأولى، ومرتفع قليلاً في السنة الثالثة (0.36+) مقارنة بالسنة الأولى، ولكنه أقل بكثير في السنة الرابعة (3.04−) مقارنة بالسنة الأولى. النتائج التي تم استخراجها بناءً على المتغيرات المتصلة كشفت ارتفاع وانخفاض بسيط بمتوسط انخفاض بلغ 1.10− في السنة.

معاملة متغير *year* كمتغير تصنيفي أو متصل يرجع للمحلل نفسه بناءً على أسباب إحصائية أو موضوعية، المتغيرات الأخرى مثل تقدير الطالب (*gpa*) فهي بوضوح متغيرات متصلة، فعندما نقوم بإدخال المتغير *gpa* ضمن المتغيرات المستقلة، فإننا نجد أنه أيضاً مرتبط بسلوك الطلبة في تناول الكحول. هذا النموذج يمزج التأثيرات التفاعلية والتي لم يتم إثبات بأنها ذات معنوية، لأن المتغيرات التصنيفية هي الوضع الافتراضي بالنسبة للأمر **anova**، أما الخيار الذي يسبق المتغير وهو **.i** فيمكن إدخاله مع المتغير *greek* أو متغير *gender*.

```
.anova drink greek gender c.gpa
```

	Number of obs = 218		R-squared = 0.2970		
	Root MSE = 5.68939		Adj R-squared = 0.2872		

Source	Partial SS	df	MS	F	Prob > F
Model	2927.03087	3	975.676958	30.14	0.0000
greek	1489.31999	1	1489.31999	46.01	0.0000
gender	405.137843	1	405.137843	12.52	0.0005
gpa	407.0089	1	407.0089	12.57	0.0005
Residual	6926.99206	214	32.3691218		
Total	9854.02294	217	45.4102439		

من هذا التحليل، يمكننا معرفة وجود علاقة ذات معنوية بين سلوك الطلبة في تناول الكحول *drink* وتقدير الطالب *gpa* وذلك عند استخدام المتغير *greek* والمتغير *gender* كمتغيرات ضابطة، ويجب ملاحظة أن

اختبارات F للمعنوية الإحصائية لا تعرض معلومات تفصيلية عن كيفية ترابط المتغيرات، ويُعتبر تحليل الانحدار أفضل من غيره للقيام بذلك.

.regress

```
      Source |       SS       df       MS              Number of obs =     218
-------------+------------------------------           F(  3,   214) =   30.14
       Model |  2927.03087     3  975.676958           Prob > F      =  0.0000
    Residual |  6926.99206   214  32.3691218           R-squared     =  0.2970
-------------+------------------------------           Adj R-squared =  0.2872
       Total |  9854.02294   217  45.4102439           Root MSE      =  5.6894

------------------------------------------------------------------------------
       drink |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1.greek |   6.547869   .9653204     6.78   0.000     4.645116    8.450623
    1.gender |   2.808418   .7938269     3.54   0.000     1.243697    4.373139
         gpa |  -3.038966   .8570168    -3.55   0.000    -4.728241    -1.34969
       _cons |   24.72871   2.539529     9.74   0.000     19.72301     29.7344
------------------------------------------------------------------------------
```

القيم المتوقعة والرسم البياني لأعمدة الخطا :

Predicted Values and Error-Bar Charts

الأمر **anova** يليه الأمر **predict** الذي يقوم بحساب القيم المتوقعة والبواقي أو الأخطأ المعيارية والإحصائية التشخيصية. أحد استخدامات مثل هذه الإحصائيات هو رسم بياني يمثل نتائج النموذج، مثل الرسم البياني لأعمدة الخطأ. لشرح ذلك سوف نعود إلى تحليل التباين ذي الاتجاه الواحد لمتغير *drink* ومتغير *year*.

.anova *drink year*

```
                  Number of obs =     243     R-squared     =  0.0609
                  Root MSE      = 6.55489     Adj R-squared =  0.0491

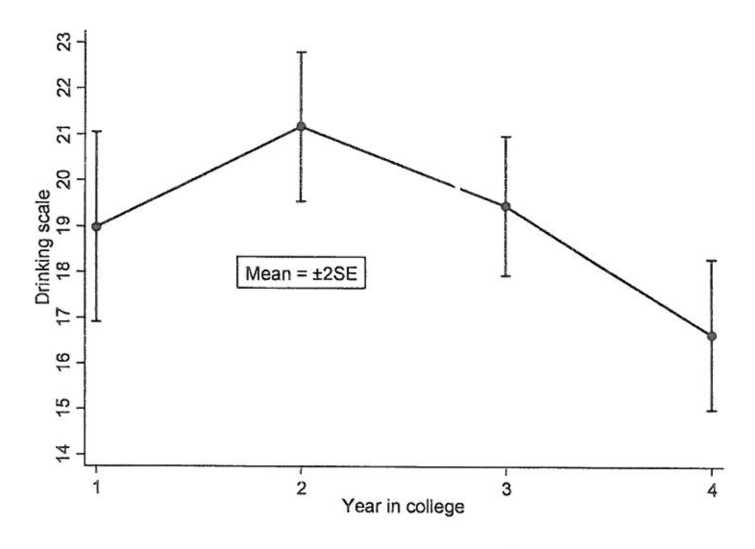
      Source |  Partial SS    df       MS           F     Prob > F
  -----------+----------------------------------------------------
       Model |  666.200518     3  222.066839       5.17     0.0018
             |
        year |  666.200518     3  222.066839       5.17     0.0018
             |
    Residual |  10269.0176   239  42.9666008
  -----------+----------------------------------------------------
       Total |  10935.2181   242  45.1868517
```

لحساب المتوسطات المتوقعة من تحليل **anova** نقــوم بطباعــة الأمــر **predict** *newvar1* حيث إن "newvar1" يمكن أن يكون أي اسم متغير تريــد معرفة متوسطاته المتوقعة، أما الأمر **predict** *newvar2*,**stdp** يقــوم بإنــشاء متغير ثان جديد يحتوي على الأخطاء المعيارية للمتوسطات المتوقعة.

```
.predict drinkmean
.predict SEdrink, stdp
```

باستخدام المتغيرين الجديدين وهما المتغير *drinkmean* والمتغير *SEdrink* يمكننا حساب 95% تقريباً من فترات الثقة، وهي عبارة عــن المتوســطات زائداً أو ناقصاً 2 الأخطاء المعيارية، الرسم البياني لأعمدة الخطأ في الشكل (3.6) تحتوي على رسم بياني في أعلاه رؤوس مدببة (**rcap**) لأعمدة الخطأ، وتم تركيب رسم بياني آخر فوقه لخط متصل (**connect**) للمتوسطات.

```
.gen drinkhi = drinkmean + 2 * SEdrink
.gen drinklo = drinkmean - 2 * SEdrink
.graph twoway rcap drinkhi drinklo year,
 color(maroon)
|| connect drinkmean year, lwidth(medthick)
   color(maroon)
|| , ylabel(14(1)23, grid gmin gmax)
   ytitle("Drinking scale")
legend(off) text(18 2 "Mean `=char(177)'2SE",
box margin(small))
```

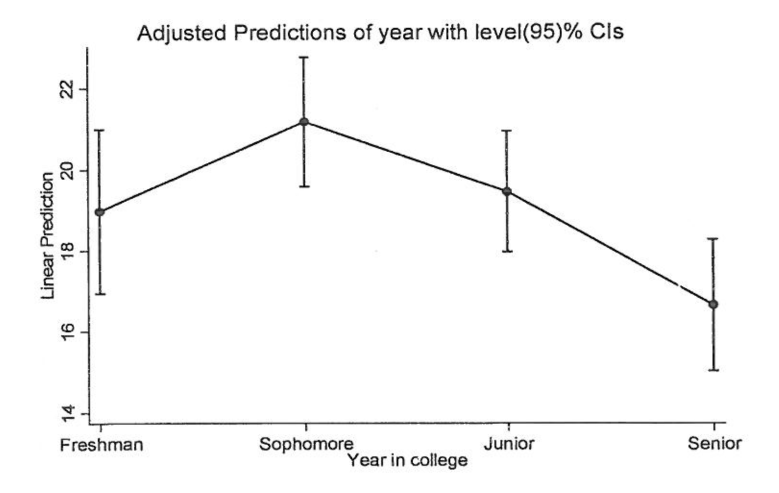


الشكل (3.6)

الشكل (3.6) يحتوي على عدد من الخيارات الأخرى لجعل الشكل أكثر وضوحاً، فالخياران **rcap** و**connect** تم إعطاؤهما نفس اللون **color(maroon)**، ومربع شرح الرسم تم إيقافه عن طريق كتابة الخيار **(legend(off))** وذلك في مقابل مربع نصّي صغير لتوضيح أن الرسم البياني يعرض "Mean ±2SE"، وعلامة الزائد أو الناقص ± عبارة عن الرمز 177 في ASCII والتي تم تمثيلها في الأمر بواسطة 'char(177)'= الشكل (16.3) في الفصل (3) يعرض مجموعة كاملة من رموز ASCII المتوافرة للاستخدام في الرسومات البيانية ببرنامج ستاتا.

الشكل (3.6) بهذه الطريقة يزودنا بمقدمة للأمر **predict** والذي له العديد من التطبيقات في النماذج الإحصائية، وهناك طريقة أخرى لرسم أعمدة الخطأ وذلك عن طريق استخدام الأمر **margins** والأمر **marginsplot**، حيث يقوم الأمر **margins** بحساب المتوسطات الحدية أو المتوسطات المتوقعة بعد أمر النموذج الإحصائي، أما الأمر **marginsplot** فيقوم بعرض بياني لكل هذه الحسابات. ففي المثال أدناه الأمر **margins** *year* يقوم بحساب قيم المتوسط للمتغير *drink* لكل سنة، ثم يقوم الأمر **marginsplot** بإنشاء رسم بياني مع فترة الثقة الخاصة به، الشكل (4.6) عبارة عن شكل واضح ولكن ليس بالإمكان تطبيق العديد من خيارات الأمر **twoway** على الأمر **marginsplot**.

```
.margins year
.marginsplot
```



الشكل (4.6)

بالنسبة لاختبار التباين العاملي ذي الاتجاهين، فإن أعمدة الخطأ تساعدنا في معاينة التأثيرات التفاعلية والرئيسة. ففي المثال أعلاه، قمنا باستخدام مقياس للسلوك العدواني *aggress* وتم اعتباره متغيرًا تابعاً. في تحليل التباين العاملي كل من المتغير *year* والمتغير *gender* وشرط التفاعل *gender#year* جميعها تعتبر مؤشرات تنبؤ، وفي ضوء العلاقة غير الخطية، فإن تأثيرات المتغير *year* يمكن مشاهدتها في الشكل (3.6) والشكل (4.6) لهذا التحليل. ويمكننا قبول التعامل الافتراضي للمتغير *year* وهي معاملته كمتغير تصنيفي بدلاً من اعتباره متغيراً متصلاً. كما أن اختبارات *F* توضح بأن متغير *gender* ومتغير *year* وشرط التفاعل *gender#year* جميعها لها تأثيرات معنوية.

```
.anova aggress gender year gender#year
```

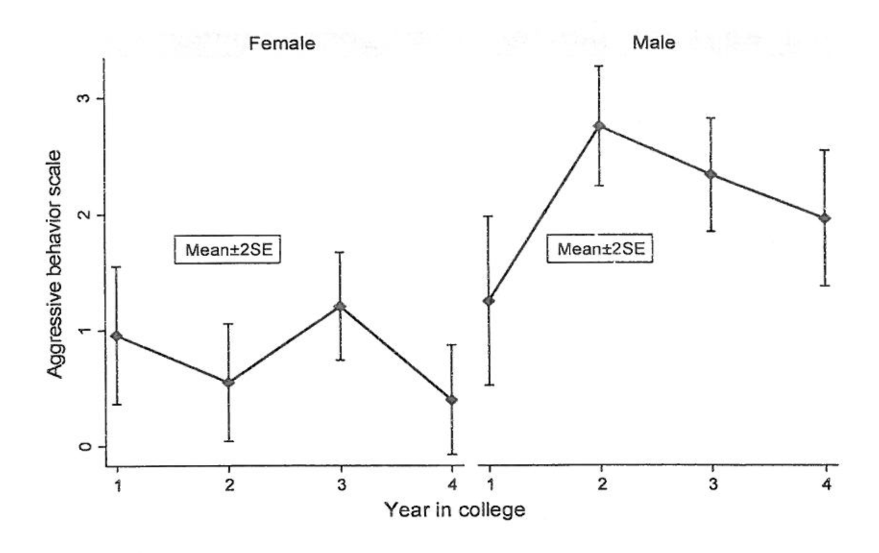
```
                     Number of obs =       243     R-squared     =  0.2503
                     Root MSE      = 1.45652     Adj R-squared =  0.2280
```

Source	Partial SS	df	MS	F	Prob > F
Model	166.482503	7	23.7832147	11.21	0.0000
gender	94.3505972	1	94.3505972	44.47	0.0000
year	19.0404045	3	6.34680149	2.99	0.0317
gender#year	24.1029759	3	8.03432529	3.79	0.0111
Residual	498.538073	235	2.12143861		
Total	665.020576	242	2.74801891		

قمنا باستخدام الأمر **predict** لحساب متغير جديد يحتوي على المتوسطات المتوقعة، واستخدمنا **predict, stdp** لحساب الأخطاء المعيارية، الحدود العليا والدنيا لفترة الثقة تساوي تقريباً زائد أو ناقص 2 الأخطاء المعيارية. لتمثيل شرط التفاعل *gender#year* بيانياً سوف نستخدم الأمر **graph** لإنشاء الشكل (5.6) مع الخيار **by(*gender*)** لرسم أشكال بيانية منفصلة تمثل الذكور والإناث. بعض الخيارات الأخرى تقوم بالتحكم في التفاصيل الثانوية الأخرى مثل العلامات بالرسم البياني (علامات ماسيّة كبيرة Diamonds) وإيقاف ظهور كل من مربع شرح الرسم البياني **legend**، ومربع الملاحظات **note**، ويمكننا أيضاً رسم مربعات صغيرة بخلفية بيضاء

حول نص "Mean±2SE" والتي يجب وضعها بعناية في داخل الرسم، بحيث لا تغطي أي بيانات داخل الرسم نفسه.

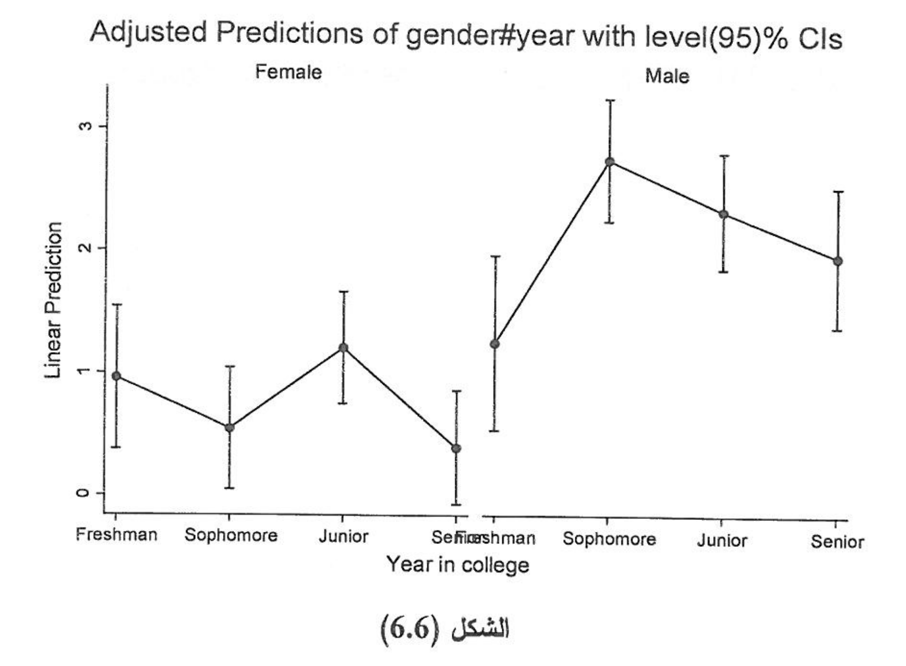
```
.predict aggmean
.predict SEagg, stdp
.gen agghi = aggmean + 2 * SEagg
.gen agglo = aggmean - 2 * SEagg
.graph twoway rcap agghi agglo year
||   connect   aggmean   year,   lwidth(medthick)
msymbol(D)
|| , by(gender, legend(off) note(""))
ytitle("Aggressive behavior scale")
text(1.7 2 "Mean `=char(177)'2SE", box
margin(small) bfcolor(white))
```



الشكل (5.6)

يمكن إنشاء رسم بياني آخر لأعمدة الخطأ بطريقة أسرع باستخدام الأمر **margins** والأمر **marginsplot**. الأمر **margins gender#year** يحسب القيم المتوقعة أو المتوسطات للمتغير *drink*، موضحاً في الرسم ذاته الجنس *gender* والسنة *year*، ثم يقوم الأمر **marginsplot, by(*gender*)** بإنشاء الرسم البياني لهذه المتوسطات مع فترات الثقة لكل جنس (الشكل 6.6).

```
.margins gender#year
.marginsplot, by(gender)
```



الشكل (6.6)

الشكل (5.6) والشكل (6.6) يضيفان تفاصيل حول الجنس، والتأثيرات التفاعلية والتي تم حسابها بواسطة الأمر **anova**. فمتوسطات الإنـــاث فـــي مقياس السلوك العدواني، شهدت بعض التقلبات مقارنة بمستوياتها المنخفضة خلال السنوات الأربع في الكلية، أما متوسطات الذكور فكانت أعلى خـــلال الفترة مع بلوغها أقصى مستوى لها في السنة الثانية مشابهةً للـــنمط الـــذي رأيناه سابقاً لسلوك الطلبة في تناول الكحول (الشكل 2.6 والشكل 3.6). لذا فإن العلاقة بين المتغير *aggress* الذي يمثل مقياساً للسلوك العدواني، ومتغير السنة *year* تختلف من الذكور للإناث. الرسومات البيانية لأعمدة الخطأ هـــي عبارة عن تكملة مرئية لجداول الأمر **anova** والأمـــر **regress**، فالرســـومات البيانية تم إنشاؤها بناءً على بيانات تلك الجداول، بينما الجداول تؤكد على أن التأثيرات ذات معنوية. وهذه الجداول تعرض تفاصيل رقمية، إلا أن استخدام الرسومات البيانية يساعد في فهم معني هذه التأثيرات.

الفصل السابع

تحليل الانحدار الخطي

Linear Regression Analysis

"ستاتا" يوفر عددًا كبيرًا من طرق تحليل الانحدار. يمكنك قراءة قائمة جزئية بالطرق المحتملة عن طريق طباعة الأمر **help regress**. هذا الفصل يركّز على الانحدار البسيط، والانحدار المتعدد، مستخدماً طريقة المربعات الصغرى العادية (OLS)، والتي يمكن القيام بها باستخدام الأمر **regress**، والأوامر الأخرى المتعلقة بهذا الأمر. الطرق التشخيصية والبيانية التي تأتي بعد تحليل الانحدار، عبارة عن امتداد لأدوات تحليل الانحدار وتساعد في تفسير النتائج، كما أنها تكشف وتتعامل مع الأمور المعقدة في التحليل. الأمر **regress** باستطاعته القيام ببعض التحليلات الأخرى غير تحليل المربعات الصغرى، وهذه التحليلات تتضمن المربعات الصغرى المرجحة. وسوف يتم شرح طرق التحليل الانحدار الأخرى في الفصل (8) والفصول اللاحقة له.

القوائم التالية تمكنك من الوصول للعمليات المطلوبة لتحليل الانحدار.

Statistics > Linear models and related > Linear regression
Statistics > Linear models and related > Regression diagnostics
Graphics > Twoway graph (scatter, line, etc.)
Statistics > Postestimation > Predictions, residuals, etc.
Statistics > Postestimation > Marginal means and predictive margins
Statistics > Postestimation > Margins plots and profile plots

هذا الفصل، يوضح بعض طرق إنشاء الرسومات البيانية لنماذج الانحدار، ويمكنك أن تجد العديد من الأمثلة في مقال *Interpreting and Visualizing Regression Models Using Stata* (Mitchell 2012).

أمثلة عن الأوامر : Example Commands

```
.regress y x
```

يقوم هذا الأمر بحساب الانحدار بطريقة المربعات الصغرى العادية (OLS) للمتغير *y* على متغير تنبؤي *x*.

```
.regress y x if ethnic == 3 & income > 50 &
 income < .
```

يحسب انحدار *y* على *x* مستخدماً جزءًا فقط من البيانات والتي فيها المتغير *ethnic* يساوي 3، والمتغير *income* أكبر من 50 (وليست هناك قيم مفقودة).

```
.predict yhat
```

إنشاء متغير جديد (تم تسميته عشوائياً باسم *yhat*) وهو يساوي القيم المتوقعة من أحدث تحليل للانحدار.

```
.predict e, resid
```

إنشاء متغير جديد (تم تسميته جزافياً باسم *e*) وهو يساوي بواقي أحدث تحليل للانحدار.

```
.predict new, cooksd
```

إنشاء متغير جديد يساوي مسافة كوك Cook's Distance ملخصاً كيف أن كل مشاهدة تؤثر في النموذج المقترح.

```
.predict new, covratio
```

إنشاء متغير جديد يساوي إحصائية بلسلي وكو وولسك Belsley, Kuh and Welsch (*COVRATIO*، وهذه الإحصائية تقيس تأثير الحالة *i*th على مصفوفة التباين – التغاير للمُعَامِلات المقدّرة.

```
.predict DFx1, dfbeta(x1)
```

إنشاء إحصائية حالة *DFBETAS* التي تقيس كيف أن كل مشاهدة تؤثر على مُعَامِلات المتغير التنبؤي *x1*، ولإنشاء مجموعة متكاملة لـ *DFBETAS*

لكل المتغيرات التنبؤية في النموذج قم بطباعة الأمر **dfbeta** بدون إضافة أي شيء آخر.

```
.predict new, dfits
```

يقوم بإنشاء إحصائيات *DFITS* التي تلخص تأثير كل مشاهدة على النموذج المقترح (هو نفس الهدف الذي تسعى إليه مسافة كوك وإحصائية ويلسك).

```
.graph twoway lfit y x || scatter y x
```

يقوم بإنشاء رسم بياني موضحاً خط الانحدار البسيط (**lfit** أو خط التطابق) مع شكل الانتشار للمتغير *y* مع المتغير *x*.

```
.graph twoway mspline yhat x || scatter y x
```

يقوم بإنشاء رسم بياني موضحاً خط الانحدار البسيط مع شكل انتشار للمتغير *y* مع المتغير *x* وذلك بواسطة خط واصل (مع مكعبات منتشرة تحت المنحنى) بين القيم المتوقعة للانحدار (في هذا المثال تم تسميتها *yhat*)، وهناك العديد من الطرق البديلة لرسم خطوط الانحدار وهذه الطرق تتضمن **mspline, mband, line, lfit, lfitci, qfit, qfitci, marginsplot** وكل طريقة لها ميزاتها وخياراتها.

```
.graph twoway scatter e yhat, yline(0)
```

يقوم برسم بياني للبواقي والقيم المتوقعة باستخدام المتغيرات *e* و *yhat* كما يمكن إنشاء نفس الرسم بطباعة الأمر **rvfplot** (البواقي) بعد تحليل الانحدار الناتج وإنشاء الرسم البياني.

```
.regress y x1 x2 x3
```

يقوم بحساب الانحدار المتعدد للمتغير *y* مع ثلاثة متغيرات تنبؤية هي *x1, x2, x3*

```
.test x1 x2
```

يقوم بحساب اختبار *F* لفرضية العدم التي تفترض أن المعامل *x1* والمعامل *x2* مساويان للصفر في آخر نموذج الانحدر.

```
.regress y x1 x2 x3, vce(robust)
```

يقوم بحساب الثقة (هيوبر/ وايت) (Huber/White) التي تقدّر الأخطاء المعيارية، انظر دليل المستخدم *User's Guide* لمزيد من التفاصيل، الخيار **vce(robust)** يعمل كذلك مع العديد من الأوامر الأخرى الملائمة للنموذج.

```
.regress y x1 x2 x3, beta
```

يقوم هذا الأمر بحساب الانحدار المتعدد، ويقوم بتضمين مُعَامِلات الانحدار المعياري (أوزان بيتا) في جدول المخرجات.

```
.correlate x1 x2 x3 y
```

يقوم بإنشاء مصفوفة ارتباطات بيرسون مستخدماً المشاهدات التي لا توجد بها قيم مفقودة في كل المتغيرات التي تم إدراجها في الأمر، عند إضافة الخيار **covariance** فسوف يتم إنشاء مصفوفة التغاير – التباين بدلاً من الارتباط.

```
.pwcorr x1 x2 x3 y, sig star(.05)
```

يقوم بإنشاء مصفوفة ارتباطات بيرسون مستخدماً الحذف الثنائي للقيم المفقودة ويعرض احتمالات اختبار *t* لفرضية العدم $H_0: \rho = 0$ لكل ارتباط، الارتباطات ذات معنوية إحصائية (في هذا المثال $p< 0.05$) سوف يتم الإشارة إليها بعلامة النجمة (*).

```
.graph matrix x1 x2 x3 y, half
```

يقوم برسم مصفوفة الانتشار، وحيث إن المتغيرات المدرجة هي نفسها التي تم استخدامها في الأمر السابق، فهذا المثال يقوم بإنشاء شكل الانتشار ويجعله منظماً بالطريقة التي قام بها الأمر **pwcorr** عند إنشاء مصفوفة الارتباط، وعند إدراج المتغير المستقل (*y*) في آخر الأمر، فإن هذا يعني إنشاء مصفوفة فيها الصف السفلي عبارة عن سلسلة لنقاط المتغير *y* مع نقاط المتغير *x*.

```
.estat hottest
```

يقوم هذا الأمر بحساب اختبار كوك وويسبرج Cook and Weisberg's test لاختلاف التباين heteroskedasticity، وإذا كان لدينا سبب للشك بأن اختلاف التباين هو دالة لمتغير تنبؤي معين *x1* فيمكننا التركيز على ذلك المتغير التنبؤي، وذلك بطباعة الأمر **estat hettest** *x1*، وللحصول على قائمة كاملة بالخيارات المتوافرة مع الأمر **regress** قم بطباعة الأمر **help regress postestimation**، توجد اختيارات مختلفة للتقدير البعدي للنماذج.

```
.estat ovtest, rhs
```

يقوم بحساب اختبار خطأ محدد انحـدار رمـزي Ramsey regression specification error test للمتغيرات المهملة، الخيار **rhs** يتطلب استخدام قوى متغيرات الطرف الأيمن للمعادلة بدلاً من القوى المتوقعة للمتغيـر y (وهـو الخيار الافتراضي).

```
.estat vif
```

يقوم بحـساب عوامـل تـضخم التبـاين لاختبـار التعـدد الخطـي multicollinearity.

```
.estat dwatson
```

يقوم بحساب اختبار دوربن واتسون Durbin-Watson للارتباط الذاتي من الدرجة الأولى في السلاسل الزمنية لبيانات (**tsset**)، الفـصل (12) يوضـح أمثلة عن هذا الاختبار وبعض الإجراءات الأخرى في السلاسل الزمنية.

```
.acprplot x1, mspline msopts(bands(7))
```

يقوم بإنشاء رسم بياني للمكونات المدمجة زائداً البواقي (يُعرف أيـضاً باسم الرسم البياني للبواقي الجزئية المدمجة) في العادة أفضل مـن الأمـر **cprplot** فـي فحـص عـدم الخطيّـة nonlinearity، الخيـار **mspline msopts(bands(7))** يقوم بإنشاء خط متصل للربط بين قيم الوسيط في سـبع نطاقات عمودية، أو بدلاً عن ذلك يمكننا إنشاء منحنى خفيف منخفض بعرض 0.5 وذلك من خلال الخيارات **lowess lsopts(bwidth(.5))**.

```
.avplot x1
```

يقوم بإنشاء رسم بياني لمتغير إضافي (يطلق عليه أيضاً اسـم انحـدار جزئي أو شكل التأثير) يعرض العلاقة بين المتغير y والمتغير $x1$ وكلاهما تم ترجيحهما للمتغيرات الأخرى x، مثل هذه الأشكال تساعد على معرفة القيم المتطرفة ونقاط التأثير.

```
.avplots
```

يقوم بإنشاء رسم بياني يتضمن صورة واحدة لكل الرسـومات البيانيـة للمتغيرات المضافة أخيراً من الأمر **anova** أو الأمر **regress**.

```
.cprplot x1
```

يقوم بإنشاء رسم بياني للمكوّن زائداً الباقي (يُعرف أيضاً باسم الرسم البياني للبواقي الجزئية) يعرض العلاقة المعدلة بين المتغير *y* والمتغير التنبؤي *x1*، ومثل هذه الأشكال تساعد في التعرّف على العلاقات غير الخطيّة في البيانات.

```
.lvr2plot
```

يقوم بإنشاء رسم بياني للتأثير مع تربيع البواقي (يُعرف أيضاً باسم شكل L-R).

```
.rvfplot
```

يقوم برسم البواقي مع القيم المتوافقة (المتوقعة) للمتغير *y*.

```
.rvpplot x1
```

يقوم برسم البواقي مع قيم المتغير التنبؤي *x1*.

```
.regress y x1 x2 i.catvar i.catvar#c.x2
```

يقوم بحساب انحدار المتغير *y* على المتغيرات التنبؤية *x1, x2* ومجموعة من المتغيرات الوهمية التي يتم إنشاؤها بشكل تلقائي لتمثل فئات المتغير *catvar* ومجموعة من شروط التفاعل التي تساوي المتغيرات الوهمية مضروبة في قياس (مستمر) المتغير *x2*، للحصول على معلومات أكثر عن هذا الأمر قم بطباعة الأمر **help fvvarlist**.

```
.stepwise, pr(.05): regress y x1 x2 x3
```

يقوم بحساب الانحدار المتدرج مستخدماً التقريب حتى تصبح جميع المتغيرات التنبؤية المتبقية ذات معنوية عند مستوى 0.05. كل المتغيرات التنبؤية المدرجة يتم إدخالها في أول تكرار، لذا فكل تكرار يقوم بحذف متغير تنبؤي واحد له أعلى مستوى ثقة *p* حتى تصبح احتمالات كل المتغيرات التنبؤية المتبقية أقل من احتمال الإبقاء عليه (pr(.05)، هناك عدة خيارات تسمح بالتقدم أو الاختيار الهرمي، الأمر **stepwise** يعمل العديد من أوامر النماذج الإحصائية الأخرى، وللحصول على قائمة بهذه الخيارات قم بطباعة الأمر **help stepwise**.

```
.regress y x1 x2 x3 [aweight = w]
```

يقوم بحساب انحدار المربعات الصغرى المرجحـــة (WLS) للمتغيــر *y* على المتغيرات *x1, x2, x3*، المتغير *w* يقوم بالاحتفاظ بــالأوزان التحليليـــة، والقيام بهذا أشبه بقيامنا بضرب كل متغير، وكل ثابت في الجذر التربيعــي للمتغير *w* ثم حساب الانحدار. الأوزان التحليلية في العادة يــتم اســتخدامها لتصحيح اختلاف التباين heteroskedasticity عندما يكون المتغير *y* والمتغير *x* عبارة عن متوسطات أو معدلات أو نسب، ويكون المتغير *w* عبارة عن عدد أفراد (مدن أو مدارس.. الخ) تمثل مجموع كل مشاهدة فــي البيانـــات، وإذا كان المتغير *y* والمتغير *x* تمثل مستويات فردية وأوزانها تشير إلــى عــدد المشاهدات المتكررة، فيمكننا استخدام الأوزان التكرارية **[fweight = *w*]**، قــم بطباعة الأمر **help survey** للحصول على معلومات عن الأوزان التي تعكس عناصر التصميم مثل العينات غير المتناسبة (انظر الفصل 4).

```
.svy: regress y x1 x2 x3
```

يقوم هذا الأمر بحساب الانحدار المرجح بالدراسة الاستقصائية للمتغير *y* على المتغيرات *x1, x2, x3* مفترضاً بأن نوع البيانات الاستقصائية تم تحديده سابقاً باستخدام الأمر **svyset** (انظر الفصل 4).

الانحدار البسيط : Simple Regression

الملف *Nations2.dta* يحتوي على بيانات عن مؤشرات التنميـــة البـــشرية للأمم المتحدة لعدد 194 دولة.

```
.use C:\data\Nations2.dta, clear
.describe country region life school chldmort
 adfert gdp
```

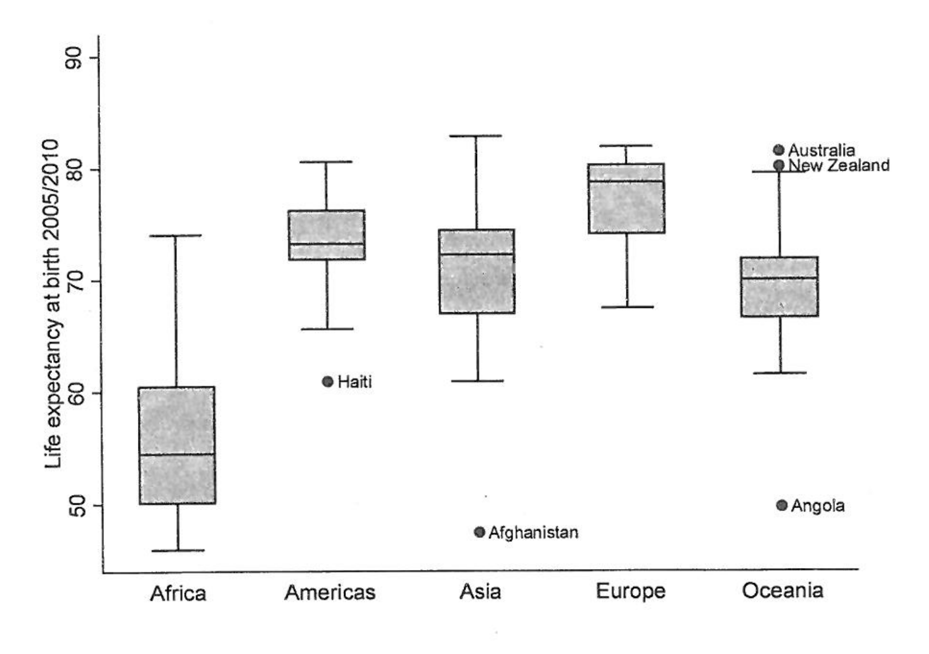
variable name	storage type	display format	value label	variable label
country	str21	%21s		Country
region	byte	%8.0g	region	Region
life	float	%9.0g		Life expectancy at birth 2005/2010
school	float	%9.0g		Mean years schooling (adults) 2005/2010
chldmort	float	%9.0g		Prob dying before age 5/1000 live births 2005/2009
adfert	float	%8.0g		Adolescent fertility: births/1000 fem 15-19, 2010
gdp	float	%9.0g		Gross domestic product per cap 2005$, 2006/2009

```
.summarize life school chldmort adfert gdp
```

Variable	Obs	Mean	Std. Dev.	Min	Max
life	194	68.7293	10.0554	45.85	82.76666
school	188	7.45922	2.959589	1.15	12.7
chldmort	193	47.65026	52.8094	2.25	209
adfert	194	51.81443	44.06612	1	207.1
gdp	179	12118.74	13942.34	279.8	74906

العمر المتوقع (*life*) يُظهِر تبايناً واضحاً من دولة لأخرى، فمثلاً الشكل (1.7) يوضح أن العمر المتوقع يبدو أنه أقل في أفريقيا عنه في الأماكن الأخرى.

```
.graph box life, over(region) marker(1,
 mlabel(country))
```



الشكل (1.7)

إلى أي مدى يمكن شرح التباين في العمر بناءً على متوسط فترة التعليم والثروة لكل فرد ومؤشرات التنمية الأخرى؟ قد نقوم بدراسة تأثيرات التعليم بواسطة حساب الانحدار البسيط للعمر المتوقع على متوسط عدد سنوات

التعليم، يمكن كتابة أمر الانحدار ببرنامج ستاتا على الشكل **regress** *yx* حيث إن *y* يمثل المتغير التابع أو المتوقع، و*x* يمثل المتغير المستقل أو التنبؤي.

```
.regress life school
```

```
      Source |       SS       df       MS              Number of obs =     188
-------------+------------------------------           F(  1,   186) =  206.34
       Model |  9846.65406     1  9846.65406           Prob > F      =  0.0000
    Residual |  8875.86926   186  47.7197272           R-squared     =  0.5259
-------------+------------------------------           Adj R-squared =  0.5234
       Total |  18722.5233   187  100.120446           Root MSE      =  6.9079

------------------------------------------------------------------------------
        life |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      school |    2.45184   .1706856    14.36   0.000     2.115112    2.788569
       _cons |   50.35941    1.36924    36.78   0.000     47.65817    53.06065
------------------------------------------------------------------------------
```

كما هو متوقع، فإن العمر المتوقع يميل ليكون أعلى في الدول التي بها عدد سنوات التعليم أعلى. من السابق لأوانه تفسير هذه النتائج عند هذه النقطة، ولكن جدول الانحدار يعطي معلومات عن علاقة إحصائية خطيّة بين العمر المتوقع *life* وعدد سنوات التعليم *school*. في الجزء الأيمن العلوي من الجدول يمكننا أن نرى نتيجة اختبار *F* والتي تم إجراؤها بناءً على مجموع التربيعات في الجزء الأعلى الأيسر بالجدول. اختبار *F* يُقيّم فرضية العدم والتي تكون جميع مُعَامِلاتها في جميع متغيرات *x* في النموذج (في مثالنا هذا هناك متغير *x* واحد وهو *school*) يساوي صفراً، إحصائية *F* 206.34 مع درجات حرية 1 و186 تشير بشكل واضح إلى رفض فرضية العدم ($p = .0000$ التي تشير إلى أربعة أرقام بعد الفاصلة العشرية، وهذا يعني أن $p < .00005$)، فعندما يكون Prop > F فهذا يعني "احتمالية أن إحصائية *F* تكون كبيرة" إذا قمنا باستخراج العديد من العينات العشوائية من المجتمع الذي تكون فيه فرضية العدم صحيحة.

في الجانب الأعلى الأيمن، يمكننا أن نرى معامل التحديد $R^2 = 0.5259$، عدد سنوات التعليم يشرح حوالي 53% من التباين في العمر المتوقع، R^2

المعدلة $R^2_a = 0.5234$ تأخذ في الاعتبار تعقيد النموذج بالمقارنة إلى تعقيد البيانات.

النصف السفلي من جدول الانحدار، يوضح النموذج المناسب نفسه، حيث نجد أن المُعَامِلات (الميل والتقاطع مع المحور الرأسي) في العمود الأول، قيمة المعامل للمتغير *school* تساوي 2.45184 والتقاطع مع المحور الأفقي (تم إدراج المُعَامِل فيcons_) بالعمود الأول وهي تساوي 50.35941 وبهذا فإن معادلة الانحدار للعمر المتوقع سوف تكون:

$$\text{المتوقع } life = 50.36 + 2.45school$$

كل سنة إضافية في عدد سنوات التعليم سوف تؤدي إلى زيادة العمر المتوقع بمقدر 2.45 سنة. هذه المعادلة تقدّر بأن العمر المتوقع هو 50.36 سنة في الدولة التي يكون فيها متوسط سنوات التعليم صفراً، هذا بالرغم من أن أقل قيمة في عدد سنوات التعليم بالبيانات الموجودة لدينا هو 1.15 سنة (وهو في دولة موزمبيق).

العمود الثاني، يعرض الأخطاء المعيارية المقدّرة للمُعَامِلات، وهذه الأخطاء يمكن استخدامها لحساب اختبارات t (الأعمدة 3–4) وفترات الثقة (الأعمدة 5–6) لكل معامل من مُعَامِلات الانحدار، إحصائيات t(المُعَامِلات قسمة أخطائها المعيارية) تختبر فرضيات العدم التي تفترض بأن مُعَامِلات المجتمع المتناظرة تساوي صفرًا، عندما تكون مستويات الثقة عند $\alpha = 0.05$ أو $\alpha = 0.001$ يمكننا رفض فرضية العدم المتعلقة بالمُعامِلين الخاصين بمتغير *school* وتقاطع المحور العمودي، لأن الاحتمالين يظهران على أنهما "0.000" (وهذا يعني أن $p = < 0.0005$)، وفي العادة فإن حسابات ستاتا تعرض فترات ثقة عند مستوى 95% ولكن نستطيع إضافة مستويات أخرى لفترات الثقة وذلك من خلال تحديدها بالخيار level()، فمثلاً لعرض فترة ثقة عند مستوى 99% قم بطباعة الأمر

```
.regress life school, level(99)
```

بعد تحديد نموذج الانحدار يمكننا إعادة عرض النتائج بطباعة الأمر **regress** فقط بدون إضافة أي متغيرات أخرى، طباعة الأمر **regress, level(90)** سوف يُعيد النتائج، وهذه المرة يعرض فترة ثقة 90%. وحيث إن بيانات الملف *Nations2.dta* التي تم استخدامها في هذا المثال لا تمثل عينة عشوائية من بعض المجتمعات في بعض الدول، فإن اختبارات الفرضيات وفترات الثقة تفتقد للتفسير الواقعي.

متوسط سنوات التعليم في بعض الدول تمتد ما بين 1.15 إلى 12.7، ماذا يمكن لمتوسط العمر المتوقع أن يفعل مع نموذج التوقّع للدول الموجودة لدينا، فمثلاً ماذا يعني أن متوسط سنوات التعليم هو 2 أو 12؟ الأمر **margins** يوفر طريقة سريعة لمراجعة المتوسطات المتوقعة مع فترات ثقتها واختبارات *z* (والتي تكون في العادة غير مهمة) وما إذا كانت هذه المتوسطات تبتعد عن الصفر، الخيار **vsquish** "التخفيض العمودي" يقلل عدد الأسطر الخالية بين الصفوف في الجدول.

```
.regress life school, level(99)
```

```
Adjusted predictions                              Number of obs   =        188
Model VCE      : OLS

Expression     : Linear prediction, predict()
1._at          : school           =           2
2._at          : school           =          12
```

	Margin	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
_at						
1	55.26309	1.059291	52.17	0.000	53.18692	57.33927
2	79.78149	.9244047	86.31	0.000	77.96969	81.59329

عندما يكون المتغير *school* يساوي 2، فإن متوسط العمر المتوقع يساوي 55.26 سنة مع فترة ثقة تتراوح بين 53.19 إلى 57.34، وعندما يكون المتغير *school* يساوي 12، فإن متوسط العمر المتوقع يساوي 79.78 سنة مع فترة ثقة تتراوح بين 79.78 إلى 81.59، ويمكننا الحصول على متوسطات العمر المتوقع لقيم المتغير *school* عند فترة ثقة سنة واحدة من 2 إلى 12 ونتائج الرسم البياني وذلك بطباعة أمرين اثنين هما:

```
.margins, at(school = (2(1)12)) vsquish
.marginsplot
```

في جدول الانحدار المصطلح cons_ يعبّر عن ثابت الانحدار، وهو في العادة يساوي واحد (إذن قيمة معامل cons_ يساوي الميل على المحور العمودي). ستاتا يقوم تلقائياً بتضمين ثابت معين ما لم نقم نحن بتغيير ذلك، والخيار **nocons** يجعل ستاتا يقوم بإيقاف الثابت، ويقوم بحساب الانحدار باستخدام المعادلة الأصلية:

```
.regress y x, nocons
```

في بعض التطبيقات، قد نحتاج إلى تحديد ثابت معين، فإذا كانت متغيرات الطرف الأيمن تتضمن ثابتاً قمنا بتحديده (اسمه *c* مثلاً) يستخدم الخيار **hascons** بدلاً من الخيار **nocons**.

```
.regress y c x, hascons
```

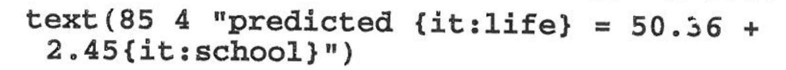
استخدام الخيار **nocons** في هذا الموضع يؤدي إلى نتائج مضللة في اختبار *F* وR^2، لمزيد من المعلومات عن الخيار **hascons** قم بالاطلاع على دليل *Base Reference Manual* أو قم بطباعة الأمر **help regress**.

الانحدر مع متغير تنبؤي واحد للحصول على خط مستقيم يتناسب بشكل كبير مع انتشار البيانات، مع تحديد "أفضل تناسب" بواسطة معيار طريقة المربعات الصغرى (OLS)، هناك طريقة سهلة لرسم هذا الخط في داخل شكل الانتشار (**twoway scatter**) مع تناسب خطي (**lfit**) في الرسم البياني نفسه، الأمر أدناه يقوم بإنشاء نسخة مبسّطة (لن يتم عرضها هنا).

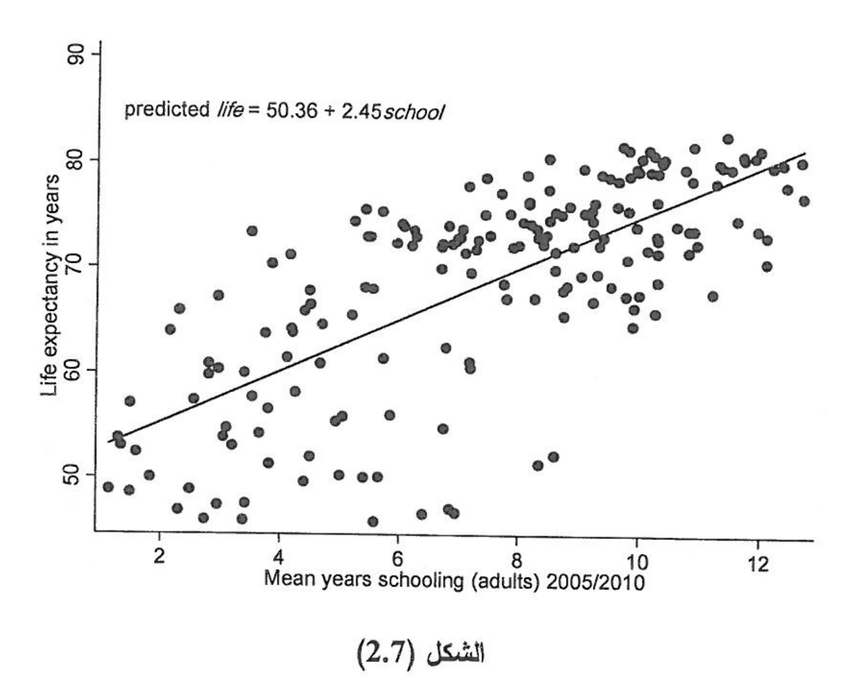
```
.graph twoway scatter life school || lfit life
 school
```

الشكل (2.7) يعرض نسخة أفضل لاغياً مربع شرح الرسم ومدرجاً معادلة الانحدار كنص، أسماء المتغيرات *life* و*school* تم كتابتها بخط مائل في الرسم البياني.

```
.graph twoway scatter life school || lfit life
 school
|| , legend(off) ytitle("Life expectancy in
 years")
```



```
text(85 4 "predicted {it:life} = 50.36 +
 2.45{it:school}")
```



الشكل (2.7)

الارتباط : Correlation

الانحدار بطريقة المربعات الصغرى (OLS) يُظهر خط التناسب المستقيم، ومُعَامِل ارتباط لحظة التوقع لبيرسون يشرح كيف يمكن لخط التناسب أن يظهر بشكل أفضل، الأمر **correlate** يقوم بحساب الارتباطات للمتغيرات المدرجة بالأمر:

```
.correlate gdp school adfert chldmort life
```

(obs=178)

	gdp	school	adfert	chldmort	life
gdp	1.0000				
school	0.5717	1.0000			
adfert	-0.5121	-0.6798	1.0000		
chldmort	-0.5160	-0.7724	0.7888	1.0000	
life	0.6062	0.7313	-0.7424	-0.9294	1.0000

الأمر correlate: يحسب الارتباطات بناءً على المشاهدات الموجودة في كل المتغيرات المدرجة بالأمر. من الجدول أعلاه، يمكننا أن نرى أن هنـــاك 178 دولة فقط من أصل 194 دولة في ملف البيانـــات *Nations2.dta* لـــديها بيانات كاملة للمتغيرات الخمسة جميعاً، هذه الدول 178 تتشابه مع مجموعة فرعية من المشاهدات التي يمكن استخدامها في النماذج المتناسبة مثل تحليل الانحدار المتعدد الذي يتضمن كل هذه المتغيرات.

المحللون لا يستخدمون الانحدار أو تقنيات متعددة المتغيـــرات، ولكـــن ربما يفضلون حساب الارتباطات بناءً على كل المشاهدات المتـــوافرة لكـــل زوج من المتغيرات، الأمر **pwcorr** (الارتباط الثنائي) يقوم بـــإجراء هـــذه الحسابات؛ كما يمكن أيضاً إجراء احتمالات اختبار *t* لفرضية العـــدم لكـــل ارتباط فردي يساوي صفرًا، في المثال أدناه الخيار **star(.05)** يطلب وضـــع نجمة (*) للارتباطات ذات معنوية إحصائية عند مستوى ثقة $\alpha = 0.05$

```
.pwcorr gdp school adfert chldmort life,
 star(.05)
```

	gdp	school	adfert	chldmort	life
gdp	1.0000				
school	0.5733*	1.0000			
adfert	-0.5171*	-0.6752*	1.0000		
chldmort	-0.5160*	-0.7727*	0.7774*	1.0000	
life	0.6112*	0.7252*	-0.7318*	-0.9236*	1.0000

هذا الأمر مهم، وسوف نستخدمه لاحقاً، ولكن إذا قمنا باستخراج العديد من العينات العشوائية من مجتمع فيه ارتباط كل المتغيرات يساوي صـــفرًا، فإن نحو 5% من ارتباطات العينات سوف تكون ذات معنوية إحصائية عند مستوى ثقة 0.05، المحللون المبتدئون الذين يختبرون العديد من الفرضيات الفردية مثل تلك التي في مصفوفة **pwcorr**، لتحديـــد الجـــزء ذي المعنويـــة الإحصائية عند مستوى ثقة 0.05 نقوم بحساب مخاطرة الحصول على الخطأ من النوع الأول عند مستوى أعلى من 0.05، هذه المشكلة يُطلق عليها أحياناً خطأ المقارنات المتعددة، الأمر **pwcorr** يعطي طريقتين لتعـــديل مـــستويات

المعنوية وأخذ مشكلة خطأ المقارنات المتعددة في الاعتبار. وهاتان الطريقتان هما بونفيروني وسيداك Bonferroni and Sidak. بالطبع طريقة سيداك ليست دقيقة، ولكن احتمالات اختبار المعنوية يمكن تعديله لعدد المقارنات التي تم إجراؤها.

```
.pwcorr gdp school adfert chldmort life, sidak
 sig star(.05)
```

	gdp	school	adfert	chldmort	life
gdp	1.0000				
school	0.5733*	1.0000			
	0.0000				
adfert	-0.5171*	-0.6752*	1.0000		
	0.0000	0.0000			
chldmort	-0.5160*	-0.7727*	0.7774*	1.0000	
	0.0000	0.0000	0.0000		
life	0.6112*	0.7252*	-0.7318*	-0.9236*	1.0000
	0.0000	0.0000	0.0000	0.0000	

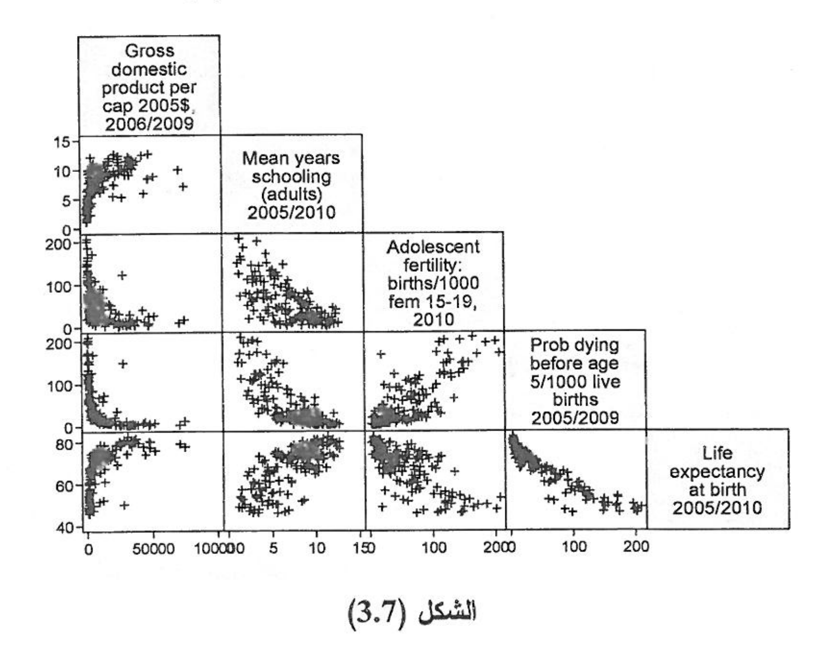
التعديلات لها تأثيرات بسيطة على الارتباطات المعتدلة والقوية، كما في الجدول أعلاه، ولكن هذا قد يكون مهماً جداً مع الارتباطات الضعيفة أو مع عدد أكثر من المتغيرات. وبشكل عام كلما زادت المتغيرات التي نقوم بحساب ارتباطاتها، فإن الاحتمالات المعدّلة سوف تزيد عن نظيراتها غير المعدّلة. انظر دليل المستخدم *Base Reference Manual* للحصول على تفاصيل أكثر عن الأمر **oneway** للمعادلات المستخدمة.

وحيث إن ارتباطات بيرسون تقيس إلى أي مدى خط الانحدار OLS متناسب، فإن مثل هذه الارتباطات تشترك مع فرضيات ونقاط ضعف OLS. وبشكل عام، فإن الارتباطات لا ينبغي تفسيرها دون النظر إلى شكل الانتشار المتعلق بها، ومصفوفات شكل الانتشار تعتبر طريقة سريعة للقيام بهذه العملية، وذلك باستخدام نفس التنظيم كمصفوفة ارتباط، الشكل (12.3) في

الفصل (3) يعرض مصفوفة شكل الانتشار المتعلقة بالأمر **pwcorr** التي تم ذكره سابقاً. الوضع الافتراضي هو قيام الأمر **graph matrix** بحذف ثنائي مثل ما يقوم به الأمر **pwcorr**، لذا فإن كل شكل انتشار صغير يعرض كل المشاهدات الموجودة فعلاً في زوج المتغيرات الموجود بالأمر.

للحصول على مصفوفة شكل انتشار تتوافق مع الأمر **correlate** أو الانحدار المتعدد من كل المشاهدات التي بها قيم مفقودة مهملة يجب علينا تعديل الأمر. إحدى طرق التعديل تتم باستثناء المشاهدات التي بها قيم مفقودة في أي متغير مدرج بالأمر، وذلك باستخدام دالة **(!missing)** (الشكل 3.7).

```
.graph matrix gdp school adfert chldmort life
if !missing(gdp,school,adfert,chldmort,life),
half msymbol(+)
```



الشكل (3.7)

الشكل (3.7) يوضح أشياء لم توضحها مصفوفة الارتباط، حيث إن العلاقات التي تتضمن GDP لكل فرد هي بوضوح علاقات غير خطية، وبالتالي فإن مُعَامِلات الارتباط أو الانحدار الخطي تعطي شرحاً غير واضح لهذه العلاقات واختباراتها المعنوية غير صالحة.

إضافة الخيار **covariance** بعد الأمر **correlate** يقوم باستخراج مصفوفة للتباين والتغاير بدلاً من الارتباطات.

```
.correlate w x y z, covariance
```

قم بطباعة الأمر أدناه بعد تحليل الانحدار، فسوف يعرض الارتباط بين المُعَامِلات المقدّرة، والتي تُستعمل أحياناً لتشخيص وجود مشكلة التعدد الخطي multicollinearity.

```
.estat vce, correlation
```

الأمر أدناه سوف يعرض المُعَامِلات المقدّرة لمصفوفة التباين – التغاير والتي تم منها استخراج الأخطاء المعيارية.

```
.estat vce
```

بالإضافة إلى ارتباطات بيرسون، فإن برنامج ستاتا يمكنه أيضاً حساب عدد من الارتباطات الترتيبية، وهذه الارتباطات يمكن استخدامها لقياس العلاقات بين المتغيرات الترتيبية أو بديل مضاد القيم المتطرفة لارتباط بيرسون لحساب المتغيرات، وللحصول على ارتباط الرتب لسبيرمان بين متغير *life* ومتغير *school* وهو مكافئ لارتباط بيرسون إذا تم وضع هذه المتغيرات في شكل ترتيبي، قم بطباعة الأمر

```
.spearman life school

      Number of obs =      188
     Spearman's rho =        0.7145

     Test of Ho: life and school are independent
         Prob > |t| =        0.0000
```

رتب الارتباطات لكندالز تاو (أ) وتاو (ب) Kendall's τ_a (tau-a) and τ_b (tau-b) يمكن إيجادها بسهولة لهذه البيانات، حتى ولو كان حجم البيانات كبيراً، فإن حساباتها سوف تكون بطيئة.

```
.ktau life school

   Number of obs =      188
 Kendall's tau-a =        0.5142
 Kendall's tau-b =        0.5149
 Kendall's score =     9039
     SE of score =      862.604   (corrected for ties)

 Test of Ho: life and school are independent
      Prob > |z| =        0.0000  (continuity corrected)
```

لغرض المقارنة، سوف نقوم بحساب ارتباط بيرسون مع قيمة *p* غير المعدلة كما يلي:

```
.pwcorr life school, sig

             |   life    school
-------------+------------------
        life |  1.0000
             |
             |
      school |  0.7252   1.0000
             |  0.0000
```

في هذا المثال قيمة الأمر **spearman** (0.71) وقيمة الأمر **pwcorr** (0.73) كلاهما تعرض ارتباطات أعلى من الأمر **ktau** (0.51)، كل هذه النتائج الثلاث تتفق بأن فرضيات العدم التي تفترض عدم وجوج علاقة يمكن رفضها.

الانحدار المتعدد : Multiple Regression

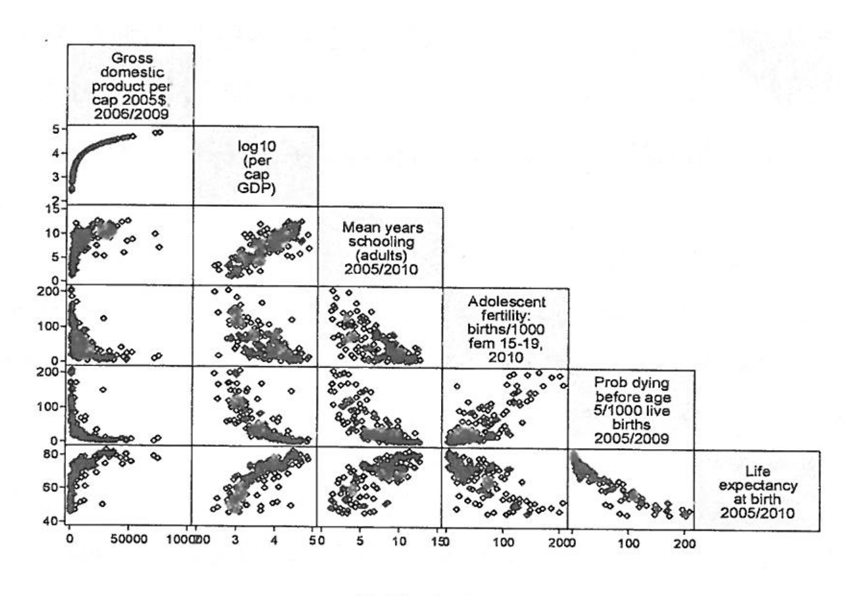
الانحدار البسيط والارتباط أظهرا بأن العمر المتوقع يرتبط بمتوسط عدد سنوات الدراسة *school*. حيث إن عدد سنوات الدراسة تشرح حوالي 52% من التباين في العمر *life*، ولكن هذه العلاقة قد تكون مفاجئة وحدثت هذه النتيجة فقط لأن المتغيرين يعكسان الوضع الاقتصادي في الدولة؟ هل متغير مدة التعليم يعتبر متغيراً مهماً عندما نتحكم في التباين بين الدول؟ هل هناك عوامل إضافية مع متغير متوسط مدة التعليم يمكن أن تشرح بنسبة أكبر من 52% التباين في العمر المتوقع؟ الانحدار المتعدد يقوم بالإجابة عن مثل هذا النوع من الأسئلة.

يمكننا إدراج متغيرات تنبؤية أخرى محتملة لمتغير *life*، وذلك بإضافة هذه المتغيرات في الأمر **regress**. فمثلاً الأمر التالي سوف يقوم بحساب انحدار العمر المتوقع على الناتج المحلي الإجمالي لكل فرد GDP ومعدل الخصوبة ومعدل وفيات الأطفال.

```
.regress life school gdp adfert chldmort
```

النتائج من الأمر السابق لم يتم إظهارها هنا، لأنها سوف تكون مضللة، ونحن نعلم من الشكل (3.7) بأن *gdp* يعرض بوضوح علاقة غير خطية مع المتغير *life* والمتغيرات الأخرى، من المفترض أن نعمل مع شكل آخر للمتغير *gdp* بحيث يعرض هذا المتغير علاقات خطية أكثر، اللوغاريتمات هي الخيار الأكثر وضوحاً وانتشاراً للتحويل، وبعد إنشاء متغير جديد يساوي لوغاريتم الأساس 10 للمتغير *gdp*، الشكل (4.7) يؤكد بأن علاقات *gdp* مع المتغيرات الأخرى تبدو أقرب لتكون خطية بالرغم من استمرارية بقاء بعض العلاقات غير الخطية.

```
.generate loggdp = log10(gdp)
.label variable loggdp "log10 (per cap GDP)"
.graph matrix gdp loggdp school adfert chldmort
 life
  if !missing(gdp,school,adfert,chldmort,life),
 half msymbol(dh)
```



الشكل (4.7)

في الفصل (8) سوف نقوم بشرح مدخل مختلف للتحويل يسمى انحدار كوكس – بوكس Box-Cox، حالياً سوف نركز على المتغير *loggdp* ونستعمله

في الأمثلة القادمة. حساب انحدار العمر المتوقع على متوسط فتـرة التعلـيم ولوغاريتم GDP ومعدل الخصوبة ومعدل الوفيات يشرح حوالي 88% مـن التباين في العمر المتوقع *life*.

```
.regress life school loggdp adfert chldmort
```

Source	SS	df	MS
Model	15545.2558	4	3886.31395
Residual	2092.93402	173	12.0978845
Total	17638.1898	177	99.6507898

Number of obs	=	178
F(4, 173)	=	321.24
Prob > F	=	0.0000
R-squared	=	0.8813
Adj R-squared	=	0.8786
Root MSE	=	3.4782

life	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
school	-.2339704	.1558288	-1.50	0.135	-.5415407	.0735999
loggdp	4.052938	.8133392	4.98	0.000	2.447592	5.658283
adfert	-.0004683	.0096839	-0.05	0.961	-.019582	.0186455
chldmort	-.1511827	.0098966	-15.28	0.000	-.1707163	-.131649
_cons	62.2544	3.114434	19.99	0.000	56.10722	68.40158

معادلة الانحدار المتعدد تكون كما يلي:

$$\text{المتوقع } life = 62.25 - 0.23school + 4.05loggdp - 0.00adfert - 0.15chldmort$$

حيث إن المعادلة أعلاه تعطينا صورة مختلفة عـن الانحـدار البـسيط السابق.

$$\text{المتوقع } life = 50.36 + 2.45school$$

عندما نتحكم في المتغيرات الثلاثة الأخرى، فإن مُعَامِل المتغير *school* يصبح سالباً وأكثر ضعفاً (0.23- ضد 2.45+) وليس ذا معنوية إحـصائية تمكننا من تمييزه عن الصفر ($t = -1.50, p = 0.135$)، معـدل الخـصوبة لـه معامل واحد على عشرين من الخطأ المعياري من الصفر، وهذا بالطبع ليس ذا دلالة إحصائية أيضاً ($t = -0.05, p = 0.961$). ومن ناحيـة أخـرى، فـإن المتغيرين *loggdp, chldmort* لهما تأثيرات جوهرية وذات معنوية إحـصائية، العمر المتوقع يميل ليكون أعلى في الدول الغنية التـي تتميـز بانخفـاض معدلات وفيات الأطفال.

مُعَامِل معدل وفيات الأطفال *chldmort* يوضح بأن العمر المتوقـع يـنخفض بمعدل 0.15 سنة عند ارتفاع معدل وفيات الأطفال بمقدار نقطة واحدة في حالـة

بقاء المتغيرات التنبؤية الأخرى على حالها، معامل *loggdp* يشير إلى زيادة العمر المتوقع بمعدل 4.05 سنة عند كل زيادة في الناتج المحلي الإجمالي لكل فرد (الناتج المحلي الإجمالي أس 10) عند ثبات كل المعدلات الأخرى، الناتج المحلي الإجمالي لكل فرد يتباين بدرجة كبيرة في البيانات الموجودة لدينا بمعدل أكبر من مائة ضعف، حيث إنه يتراوح من 279.8 دولار/ للفرد (جمهورية الكونغو الديمقراطية) إلى 74,906 دولار/ للفرد (قطر).

الأربعة متغيرات التنبؤية معاً تشرح حوالي 88% من التباين في العمر المتوقع (R^2_a=0.8786)، وتعتبر R^2_a المعدلة إحصائية مختصرة مفضلة في حالة الانحدار المتعدد، لأنها من غير المحتمل أن تعدل R^2(R^2=0.8813)، كما أن R^2_a تفرض جزاء عند إنشاء نماذج معقدة جداً، R^2 دائماً سوف تزداد عندما نقوم بإضافة متغيرات تنبؤية أكثر، ولكن R^2_a ربما لا تسلك نفس المسلك.

التأثير القريب من الصفر للمتغير *adfert* وضعف تأثيره، والتأثير غير المعنوي للمتغير *school* تشير إلى أن الأربعة متغيرات التنبؤية هي عبارة عن تعقيد غير ضروري. إن إضافة متغيرات تنبؤية لا علاقة لها بالنموذج قد تؤدي إلى ارتفاع الأخطاء المعيارية للمتغيرات التنبؤية الأخرى مؤدياً إلى انخفاض دقة التقديرات لتلك التأثيرات، يمكن الحصول على نموذج فعّال ومختصر، وذلك باستبعاد المتغيرات التنبؤية التي لم تكن ذات معنوية وذلك باستبعاد واحد في كل مرة، أولاً سوف نقوم باستبعاد المتغير *adfert*.

```
.regress life school loggdp chldmort

      Source |       SS       df       MS              Number of obs =     178
-------------+------------------------------           F(  3,   174) =  430.79
       Model |  15545.2275     3  5181.7425            Prob > F      =  0.0000
    Residual |   2092.9623   174  12.028519            R-squared     =  0.8813
-------------+------------------------------           Adj R-squared =  0.8793
       Total |  17638.1898   177  99.6507898           Root MSE      =  3.4682

------------------------------------------------------------------------------
        life |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      school |   -.233002   .1540928    -1.51   0.132    -.5371337    .0711297
      loggdp |   4.056047    .808465     5.02   0.000     2.460387    5.651708
    chldmort |  -.1514263    .008493   -17.83   0.000     -.168189   -.1346637
       _cons |   62.22201   3.032798    20.52   0.000      56.2362    68.20782
------------------------------------------------------------------------------
```

ثم نقوم باستبعاد متغير *school*.

```
.regress life loggdp chldmort
```

```
      Source |       SS       df       MS              Number of obs =     178
-------------+------------------------------           F(  2,   175) =  640.33
       Model |  15517.7253     2  7758.86267           Prob > F      =  0.0000
    Residual |  2120.46446   175  12.1169398           R-squared     =  0.8798
-------------+------------------------------           Adj R-squared =  0.8784
       Total |  17638.1898   177  99.6507898           Root MSE      =  3.4809

------------------------------------------------------------------------------
        life |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      loggdp |   3.510749   .7262322     4.83   0.000     2.077448     4.94405
    chldmort |  -.1457805   .0076563   -19.04   0.000     -.160891     -.13067
       _cons |   62.28614   3.043627    20.46   0.000      56.2792    68.29308
------------------------------------------------------------------------------
```

انتهينا بنموذج يحتوي على متغيرين تنبؤيين اثنين، وأخطاء معيارية منخفضة. وعملياً فإن قيمة R^2_a المعدلة هي نفسها (0.8784) عند وجود متغيرين تنبؤيين، و0.8786 عند وجود أربعة متغيرات) وقيمة مُعَامِلات المتغير *loggdp* أصبحت أقل بنسبة بسيطة، بينما قيمة معاملات المتغير *chldmort* استمرت ثابتة تقريباً.

$$\text{المتوقع } life = 62.29 + 3.51 loggdp - 0.15 chldmort$$

يمكننا حساب القيم المتوقعة لأي مجموعة من قيم المتغيرات *loggdp* و*chldmort* بواسطة تعويض هذه القيم في معادلة الانحدار، الأمر **margins** يقوم بحساب المتوسطات المتوقعة (تسمى أيضاً المتوسطات المعدلة) للمتغير التابع عند قيم معينة لمتغير مستقل واحد أو أكثر، فمثلاً لمعرفة متوسط العمر المتوقع و*loggdp* المعدلة عندما تكون قيم المتغير *chldmort* هي 2، 100، 200.

```
.margins, at(chldmort = (2 100 200)) vsquish
```

```
Predictive margins                                Number of obs   =        178
Model VCE    : OLS

Expression   : Linear prediction, predict()
1._at        : chldmort        =           2
2._at        : chldmort        =         100
3._at        : chldmort        =         200

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   75.23421   .4408642   170.65   0.000     74.37013    76.09828
          2  |   60.94772   .4733418   128.76   0.000     60.01998    61.87545
          3  |   46.36966   1.189535    38.98   0.000     44.03822    48.70111
------------------------------------------------------------------------------
```

جدول الأمر **margins** السابق، يوضح بأن المتوسـط المقـدّر للعمـر المتوقع عندما تكون *chldmort* = 2 خلال جميع القيم المشاهدة للمتغير *loggdp* يساوي 75.23، وبشكل مشابه، فإن المتوسط المقدّر للعمر المتوقـع عنـدما تكون *chldmort* = 200 يساوي 46.37، وإذا قمنا بإدخال الخيار **atmeans** فإن المتغير *loggdp* سوف يساوي قيمة متوسطة معطياً نتائج متساوية فـي هـذا المثال، والمخرجات سوف يتم توصيفها بأنها "التوقعات المعدلـة" "adjusted predictions" بدلاً من "الهوامش التنبؤية" "predictive margins".

يمكننا أن نقوم بحساب المتوسطات المتوقعة عند قيم محددة للمتغيـرات *chldmort* و*loggdp*. الأوامر التالية تحسب المتوسطات عند ست مجموعـات من القيم، عندما يكون المتغير *chldmort* يساوي 2 أو 100 أو 200، وعندما يكون المتغير *loggdp* يساوي 2.5 أو 4.5.

```
.margins, at(chldmort = (2 100 200) loggdp =
 (2.5 4.5)) vsquish
```

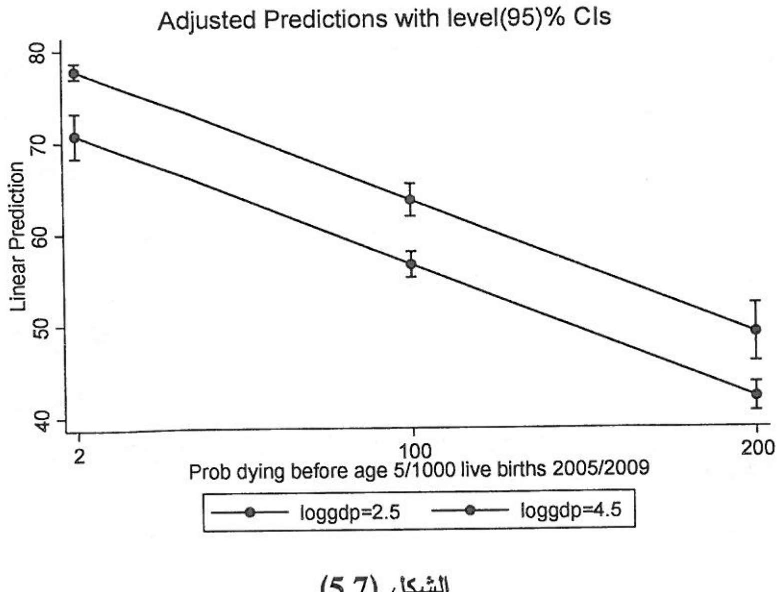
```
Adjusted predictions                              Number of obs   =        178
Model VCE      : OLS

Expression     : Linear prediction, predict()
1._at          : loggdp          =         2.5
                 chldmort        =           2
2._at          : loggdp          =         2.5
                 chldmort        =         100
3._at          : loggdp          =         2.5
                 chldmort        =         200
4._at          : loggdp          =         4.5
                 chldmort        =           2
5._at          : loggdp          =         4.5
                 chldmort        =         100
6._at          : loggdp          =         4.5
                 chldmort        =         200
```

	Margin	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
_at						
1	70.77145	1.244946	56.85	0.000	68.3314	73.2115
2	56.48496	.718987	78.56	0.000	55.07577	57.89415
3	41.90691	.7896567	53.07	0.000	40.35921	43.45461
4	77.79295	.4312218	180.40	0.000	76.94777	78.63813
5	63.50646	.9082472	69.92	0.000	61.72633	65.28659
6	48.92841	1.624056	30.13	0.000	45.74532	52.1115

الأمر أدناه **marginsplot** يقوم بإنشاء رسم بياني لنتائج الهوامش **margins** في الشكل (5.7).

```
.marginsplot
```



الشكل (5.7)

للحصول على معاملات الانحدار المعيارية (أوزان بيتا) مع الانحدار نقوم بإضافة الخيار **beta**، المُعَامِلات المعيارية هي ما يمكن أن نجده في الانحدار عندما تكون جميع المتغيرات قد تم تحويلها إلى نتائج قياسية والتي متوسطها الحسابي يساوي صفرًا وانحرافها المعياري يساوي واحدًا.

```
.regress life loggdp chldmort, beta
```

Source	SS	df	MS
Model	15517.7253	2	7758.86267
Residual	2120.46446	175	12.1169398
Total	17638.1898	177	99.6507898

Number of obs	=	178
F(2, 175)	=	640.33
Prob > F	=	0.0000
R-squared	=	0.8798
Adj R-squared	=	0.8784
Root MSE	=	3.4809

life	Coef.	Std. Err.	t	P>|t|	Beta
loggdp	3.510749	.7262322	4.83	0.000	.1974935
chldmort	-.1457805	.0076563	-19.04	0.000	-.7778774
_cons	62.28614	3.043627	20.46	0.000	.

معادلة الانحدار المعيارية سوف تكون على الشكل التالي:

$$\text{المتوقع } life^* = 0.197loggdp^* - 0.778chldmort^*$$

حيث إن $life^*$ و $loggdp^*$ و $chldmort^*$ تشير إلى أن هذه المتغيرات في شكل معياري، فمثلاً يمكننا تفسير المعامل المعياري للمتغير *chldmort* كما يلي:

$b_2^* = -0.778$ يقدّر بأن العمر المتوقع *life* ينخفض بمقدار 0.778 انحراف معياري مع كل زيادة بمقدار 1 في الانحراف المعياري لمعدل وفيات الأطفال *chldmort* إذا لم يتغير الناتج المحلي الإجمالي *loggdp*.

اختبارات F و t و R^2 والخصائص الأخرى للانحدار تبقى كما هي.

اختبارات الفرضيات : Hypothesis Tests

هناك نوعان من اختبارات الفرضيات يظهران في جداول مخرجات الأمر **regress**، الفرضيات تبدأ باعتبار أن المشاهدات الموجودة في العينة لدينا تم سحبها عشوائياً وبشكل مستقل من مجتمع كبير وغير محدود.

1- اختبار F الشامل: إحصائية F في أعلى اليمين في جدول الانحدار تقيّم فرضية العدم في المجتمع، والمُعَامِلات لكل المتغيرات في النموذج تساوي صفرًا.

2- اختبارات t الفردية: وتظهر في العمودين الثالث والرابع في جدول الارتباط، وتحتوي على اختبارات t لكل مُعامِل انحدار على حدة، وهي تقيّم فرضيات العدم في المجتمع التي تفترض بأن المُعامِلات لكل متغير معين تساوي صفرًا.

احتمالات اختبار t ذات الحدين، واختبارات الجانب الأول تقوم بقسمة قيم p بالمنتصف.

بالإضافة إلى أن اختبارات F و t المعيارية يمكن لبرنامج ستاتا حساب اختبارات F لفرضيات يحددها المستخدم، الأمر **test** يشير إلى النماذج المناسبة الأخيرة مثل **anova** أو **regress**؛ بالعودة إلى مثالنا السابق الخاص بانحدار أربعة متغيرات تنبؤية، بافتراض أننا نريد اختبار فرضية العدم التي تفترض بأن المتغيرين *adfert* و *chldmort* (يتم أخذهما بالاعتبار معاً) ليس لهما تأثير.

```
.regress life school loggdp adfert chldmort
```

```
      Source |       SS       df       MS              Number of obs =     178
-------------+------------------------------           F(  4,   173) =  321.24
       Model |  15545.2558     4  3886.31395           Prob > F      =  0.0000
    Residual |  2092.93402   173  12.0978845           R-squared     =  0.8813
-------------+------------------------------           Adj R-squared =  0.8786
       Total |  17638.1898   177  99.6507898           Root MSE      =  3.4782

------------------------------------------------------------------------------
        life |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      school |  -.2339704   .1558288    -1.50   0.135    -.5415407    .0735999
      loggdp |   4.052938   .8133392     4.98   0.000     2.447592    5.658283
      adfert |  -.0004683   .0096839    -0.05   0.961     -.019582    .0186455
    chldmort |  -.1511827   .0098966   -15.28   0.000    -.1707163    -.131649
       _cons |    62.2544   3.114434    19.99   0.000     56.10722    68.40158
------------------------------------------------------------------------------
```

```
.test adfert chldmort

 ( 1)  adfert = 0
 ( 2)  chldmort = 0

       F(  2,   173) =  158.03
            Prob > F =    0.0000
```

بينما فرضيات العدم الفردية تحدد اتجاهات معاكسة (تأثير *chldmort* ذو معنوية إحصائية أما *adfert* ليس كذلك) الفرضية مجتمعة التي تفترض بـأن مُعَاملات المتغيرات *chldmort* و *adfert* كلاهمـا يساوي صفرًا يمكن رفضه ($p< 0.00005$)، تطبيق مثل هذه الاختبارات علـى مجموعـة فرعيـة مـن المعاملات تعتبر مفيدة عندما يكون لدينا مجموعة من المتغيـرات التنبؤيـة الوهمية، أو عندما تكون تقديرات المعاملات الفردية تُظهِر عـدم مـصداقية نتيجة التعدد الخطي multicollinearity.

الأمر **test** يمكنه تكرار اختبار *F* الشامل.

```
.test school loggdp adfert chldmort

 ( 1)  school = 0
 ( 2)  loggdp = 0
 ( 3)  adfert = 0
 ( 4)  chldmort = 0

       F(  4,   173) =  321.24
            Prob > F =    0.0000
```

كما يمكن للأمر **test** تكرار اختبارات المُعَامِلات الفردية، فمثلاً بخصوص معامل المتغير *school* يمكن الحصول على إحصائية *F* عن طريق الأمر **test** وهو يساوي مربع إحصائية *t* في جدول الانحدار $(-1.50)^2 = 2.25$ وسوف يُنتج نفس قيمة *P*.

```
.test school

 ( 1)  school = 0

       F(  1,   173) =    2.25
            Prob > F =    0.1351
```

تطبيقات الأمر **test** تعتبر أكثر فائدة عند القيام بأعمال متقدمة (بالرغم من أنها عديمة الفائدة بالنسبة للعمر المتوقع في مثالنا هذا) تتضمن التالي:

1- اختبار ما إذا كان مُعَامِل ما يساوي ثابتاً معيناً، فمثلاً لاختبار فرضية العدم التي تقول بأن معامل المتغير *school* يساوي 1 $(H_0 : \beta_1 = 0)$ نقوم بطباعة الأمر:

```
.test school = 1
```

2- اختبار ما إذا كان مُعَامِلان متساويين، فمثلاً الأمر أدناه يقيّم فرضية العدم $H_0 : \beta_2 = \beta_3$.

```
.test loggdp = adfert
```

3- أخيراً الأمر **test** يستطيع تفهّم بعض التعبيرات الجبريّة، حيث يمكننا طلب شيء ما مثل اختبار $H_0\ \beta_2=(\beta_3+\beta_4)/100$.

```
.test school = (loggdp + adfert)/100
```

لمزيد من المعلومات والأمثلة حول الأمر **test** قم بطباعة الأمر **help test**

المتغيرات الوهمية : Dummy Variables

المتغيرات النوعية يمكن أن تصبح متغيرات تنبؤية في تحليل الانحدار عندما يتم التعبير عنها على شكل رقم واحد أو أكثر {0,1} فهي ثنائيات تسمى متغيرات وهمية. فعلى سبيل المثال، كنا قد لاحظنا وجود اختلاف

كبير بين مناطق العالم بالنسبة لمتوسط العمر المتوقع (الشكل 1.7)، المتغير النوعي *region* يساوي القيم من 1 (أفريقيا) إلى 5 (الجزر الأستوائية بالمحيط الهادي) والتي يمكن إعادة التعبير عن قيم هذا المتغير كمجموعة من خمسة متغيرات وهمية {0,1}، الأمر **tabulate** يوفر طريقة تلقائية للقيام بذلك حيث يقوم بإنشاء متغير وهمي واحد لكل فئة للمتغيرات المدرجة في الأمر عند القيام بإدراج الخيار **gen** (إنشاء). في المثال أدناه فإن المتغيرات الوهمية التي تم إنشاؤها تم تسميتها بأسماء من *reg1* إلى *reg5*. حيث إن *reg1* يساوي 1 للدول الأفريقية وصفر لبقية الدول، *reg2* يساوي 1 لدول أمريكا وصفر لبقية الدول، وهكذا لبقية المتغيرات.

```
.tabulate region, gen(reg)
```

Region	Freq.	Percent	Cum.
Africa	52	26.80	26.80
Americas	35	18.04	44.85
Asia	49	25.26	70.10
Europe	43	22.16	92.27
Oceania	15	7.73	100.00
Total	194	100.00	

```
.describe reg*
```

variable name	storage type	display format	value label	variable label
region	byte	%8.0g	region	Region
reg1	byte	%8.0g		region==Africa
reg2	byte	%8.0g		region==Americas
reg3	byte	%8.0g		region==Asia
reg4	byte	%8.0g		region==Europe
reg5	byte	%8.0g		region==Oceania

```
.label values reg1 reg1
.label define reg1 0 "others" 1 "Africa"
.tabulate reg1 region
```

region==Africa	Region					
	Africa	Americas	Asia	Europe	Oceania	Total
others	0	35	49	43	15	142
Africa	52	0	0	0	0	52
Total	52	35	49	43	15	194

انحدار المتغير *life* على متغير وهمي واحد *reg1* (أفريقيا) هو معادل للقيام باختبار *t* لعينتين اثنتين لمعرفة ما إذا كان متوسط المتغير *life* هو نفسه بالنسبة لفئات *reg1*؛ وهل متوسط العمر المتوقع يختلف اختلافاً معنوياً بالنسبة لأفريقيا مقارنةً بقارات العالم الأخرى؟

```
.ttest life, by(reg1)

Two-sample t test with equal variances

   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
  others |     142    73.21115    .5068244     6.03951    72.20919    74.21311
  Africa |      52    56.49038    1.185937    8.551912    54.10952    58.87125
---------+--------------------------------------------------------------------
combined |     194     68.7293    .7219359     10.0554     67.3054    70.15319
---------+--------------------------------------------------------------------
    diff |            16.72077    1.101891                 14.5474    18.89413
------------------------------------------------------------------------------
    diff = mean(others) - mean(Africa)                            t =  15.1746
Ho: diff = 0                                     degrees of freedom =      192

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000

.regress life reg1

      Source |       SS       df       MS              Number of obs =     194
-------------+------------------------------           F(  1,   192) =  230.27
       Model |  10641.4858     1  10641.4858           Prob > F      =  0.0000
    Residual |  8872.96636   192  46.2133664           R-squared     =  0.5453
-------------+------------------------------           Adj R-squared =  0.5429
       Total |  19514.4521   193  101.111151           Root MSE      =   6.798

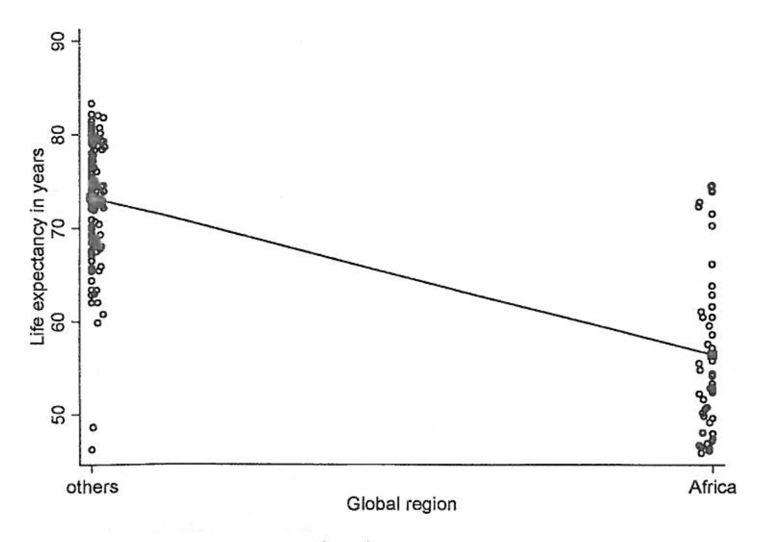
------------------------------------------------------------------------------
        life |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        reg1 |  -16.72077   1.101891   -15.17   0.000    -18.89413    -14.5474
       _cons |   73.21115    .570479   128.33   0.000     72.08594    74.33636
------------------------------------------------------------------------------
```

اختبار *t* يؤكد بأن 16.72 سنة اختلاف بين متوسطات أفريقيا (56.49) ومناطق العالم الأخرى (73.21) هو اختلاف ذو معنوية إحصائية ($t = 15.17, p = 0.000$) حصلنا على نفس النتائج من انحدار المتغير الوهمي ($t = -15.17, p = 0.000$) حيث إن مُعَامِل *reg1*($b_1 = -16.72$) كما أنه يشير إلى

أن متوسط العمر المتوقع 16.72 سنة أقل في أفريقيا عنـــه في مناطق العالم الأخرى ($b_0 = 73.21$).

الشكل (6.7) يوضح بيانياً انحدار المتغير الوهمي، كل خيارات البيانات تم تمثيلها بخط يمتد عبر مجموعتين أفقيتين عند المتغير *reg1* = 1 (أفريقيـــا) و*reg1* = 0 (غير ذلك)، لتوضيح هذه النقطة بيانياً هذا المثال يستخدم الخيـــار **jitter(5)** والذي يقوم بإضافة حجم صغير لكل نقطة في الرسم البياني بحيث تظهر على شكل دائرة صغيرة، الخيار **jitter()** لا يؤثر على خط الانحـــدار وهو يربط متوسط المتغير *life* عندما تكون *reg1* = 0(73.21) مـــع متوســـط المتغير *life* عندما *reg1* = 0(56.49) كلا المتوسطين أو القيم المتوقعـــة يـــتم رسمها باستخدام مربعات مظللة، الفرق بين المتوسطين يساوي ميل الانحدار 16.72− سنة، يجب ملاحظة أن القيم 0 و1 للمتغير *reg1* تم إعادة توصيفها بالرسم البياني باستخدام الخيار **xlabel()** بالأمر **graph**.

```
.predict lifehat
.graph twoway scatter life lifehat reg1,
msymbol(oh S) jitter(5)
 || lfit life reg1
 || , legend(off) xlabel(0 "others" 1 "Africa")
xtitle("Global region") ytitle("Life expectancy
in years")
```



الشكل (6.7)

مناطق العالم الخمس تم التعبير عنها بخمسة متغيرات وهمية، ولكن ليس من المحتمل أن يتم تضمين كل المتغيرات الخمسة في نموذج انحدار واحد بسبب الارتباط الخطي المتعدد multicollinearity وهي أن قيم أي أربعة متغيرات من هذه المتغيرات الوهمية لها القدرة على تحديد قيم المتغير الخامس، وبالتالي يمكننا تمثيل كل المعلومات للفئة *k* لمتغير نوعي ما من خلال متغيرات وهمية *k-1*. فمثلاً وكما رأينا سابقاً بخصوص الناتج المحلي الإجمالي لكل فرد (في الشكل اللوغاريتمي *loggdp*) ومعدل وفيات الأطفال (*chldmort*) معاً يوضحان حوالي 88% من التباين في العمر المتوقع، تضمين أربعة متغيرات وهمية لمناطق العالم 1–4 يزيد من هذه النسبة لتصبح حوالي 89% (R^2_a=0.8872).

```
.regress life reg1 reg2 reg3 reg4 loggdp
 chldmort
```

Source	SS	df	MS
Model	15715.8742	6	2619.31237
Residual	1922.31561	171	11.2416118
Total	17638.1898	177	99.6507898

Number of obs	=	178
F(6, 171)	=	233.00
Prob > F	=	0.0000
R-squared	=	0.8910
Adj R-squared	=	0.8872
Root MSE	=	3.3529

life	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
reg1	-2.150794	1.211918	-1.77	0.078	-4.543041	.2414527
reg2	1.486722	1.176501	1.26	0.208	-.8356127	3.809057
reg3	.9838334	1.129945	0.87	0.385	-1.246603	3.21427
reg4	1.455465	1.199846	1.21	0.227	-.9129513	3.823882
loggdp	3.239467	.7342834	4.41	0.000	1.79004	4.688894
chldmort	-.1270253	.0086971	-14.61	0.000	-.1441928	-.1098578
_cons	62.16206	3.10137	20.04	0.000	56.04016	68.28396

كل المتغيرات الوهمية الخاصة بمناطق العالم ليست ذات معنوية إحصائية عندما قمنا بإدخالها جميعاً وقمنا بالتحكم في المتغيرات *loggdp* و*chldmort*؛ المُعَاملات ليست ذات معنوية إحصائية تشير إلى أن نموذج أصغر قد يكون أكثر فائدة، ويعطي صورة أكثر وضوحاً للتأثيرات المهمة، الخطوة الأولى نحو تقليص النموذج تتضمن استبعاد المتغير *reg3* وهو أضعف متغير تنبؤي، النتائج أدناه توضح أنها أفضل (R^2_a=0.8873) وتعطي

تقديرات أكثر دقة (أخطاء معيارية أقل) لتأثيرات المناطق الأخرى، مُعَامِل المتغير *reg1* الآن يظهر ذو معنوية إحصائية.

```
.regress life reg1 reg2 reg4 loggdp chldmort

      Source |       SS       df       MS              Number of obs =     178
-------------+------------------------------           F(  5,   172) =  279.84
       Model |  15707.3519     5  3141.47038           Prob > F      =  0.0000
    Residual |  1930.83792   172  11.2258018           R-squared     =  0.8905
-------------+------------------------------           Adj R-squared =  0.8873
       Total |  17638.1898   177  99.6507898           Root MSE      =  3.3505

------------------------------------------------------------------------------
        life |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        reg1 |  -2.927382   .8199249    -3.57   0.000    -4.545793   -1.308972
        reg2 |   .6920922   .7419319     0.93   0.352    -.7723717    2.156556
        reg4 |   .6487658   .7618415     0.85   0.396    -.8549968    2.152528
      loggdp |   3.273944   .7326992     4.47   0.000     1.827705    4.720184
    chldmort |  -.1269767   .0086908   -14.61   0.000    -.1441311   -.1098224
       _cons |   62.82061    3.00561    20.90   0.000     56.88798    68.75324
------------------------------------------------------------------------------
```

الخطوة التالية، استبعاد المتغير *reg4* ثم أخيراً *reg2*، النتائج في النموذج المصغّر مازالت توضح 89% من التباين في العمر المتوقع (R^2_a=0.8879) ولكن مع ثلاثة متغيرات تنبؤية فقط.

```
.regress life reg1 loggdp chldmort

      Source |       SS       df       MS              Number of obs =     178
-------------+------------------------------           F(  3,   174) =  468.34
       Model |  15694.5388     3  5231.51293           Prob > F      =  0.0000
    Residual |  1943.65102   174  11.1704082           R-squared     =  0.8898
-------------+------------------------------           Adj R-squared =  0.8879
       Total |  17638.1898   177  99.6507898           Root MSE      =  3.3422

------------------------------------------------------------------------------
        life |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        reg1 |  -3.143763   .7901811    -3.98   0.000    -4.703336   -1.584189
      loggdp |   3.414611   .6977087     4.89   0.000     2.037549    4.791672
    chldmort |  -.1277141   .0086406   -14.78   0.000     -.144768   -.1106603
       _cons |   62.65707   2.923818    21.43   0.000     56.88636    68.42779
------------------------------------------------------------------------------
```

من هذا التحليل الإحصائي، يمكننا أن نستنتج أن الاختلافات في العمر المتوقع بين مناطق العالم المختلفة كانت بسبب التباين في الصحة ومعدل وفيات الأطفال، ولكن في أفريقيا هناك ظروف أخرى لها تأثيرها (مثل الحروب) التي تقلل من العمر المتوقع.

التأثيرات التفاعلية : Interaction Effects

الجزء السابق شرح ما يمكن أن يسمى "المتغيرات الوهمية التقاطعيــة"، لأن مُعَامِلاتها تزداد لتؤثر في تقاطع معادلة الانحدار مع المحور العمــودي مقارنة بين مجموعات 0 و1. هناك استخدام آخر للمتغيرات الوهمية وهــي إنشاء شرط تفاعلي يسمى "ميل المتغيرات الوهمية" وذلك بــضرب المتغيــر الوهمي في قياس المتغير. في هذا الجزء سوف نستمر في اســتخدام ملــف بيانات *Nations2.dta* ولكن نأخذ في الاعتبار بعض المتغيرات المختلفة، وهي انبعاث ثاني أكسيد الكربون لكل فرد (*co2*)، نسبة عدد السكان الذين يعيشون في المناطق الحضرية (*urban*)، والمتغير الوهمي الرابع *reg4* الذي يساوي 1 للدول الأوروبية و0 غير ذلك، نبدأ بتوصيف قــيم المتغيــر *reg4* وحــساب لوغاريتم للمتغير *co2* بسبب الالتواء الموجب الكبير.

```
.label values reg4 reg4
.label define reg4 0 "others" 1 "Europe"
.generate logco2 = log10(co2)
.label variable logco2 "log10(per cap CO2)"
.describe urban reg4 co2 logco2
```

variable name	storage type	display format	value label	variable label
urban	float	%9.0g		Percent population urban 2005/2010
reg4	byte	%8.0g	reg4	region==Europe
co2	float	%9.0g		Tons of CO2 emitted per cap 2005/2006
logco2	float	%9.0g		log10(per cap CO2)

نقوم بإنشاء مصطلح تفاعلي أو ميل متغير وهمي يُسمى *urb_reg4* وذلك بضرب المتغير الوهمي *reg4* في قياس المتغير *urban*، المتغير الناتج وهــو *urb_reg4* يساوي *urban* للدول الأوروبية وصفر لدول العالم الأخرى.

```
.generate urb_reg4 = urban * reg4
.label variable urb_reg4 "interaction
urban*reg4 (Europe)"
```

انحدار المتغير *logco2* على المتغيــر *urban* والمتغيــر *reg4* وشــرط التفاعل *urb_reg4* يعطي نموذجاً يختبر ما إذا كان تقاطع المحور العمودي *y*

أو الميل المتعلق بالمتغير *logco2* على المتغير *urban* ربما يختلف لدول أوروبا عن باقي دول العالم.

```
.regress logco2 urban reg4 urb_reg4
```

Source	SS	df	MS
Model	55.8882644	3	18.6294215
Residual	46.2772694	181	.255675521
Total	102.165534	184	.555247466

Number of obs =	185
F(3, 181) =	72.86
Prob > F =	0.0000
R-squared =	0.5470
Adj R-squared =	0.5395
Root MSE =	.50564

logco2	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
urban	.0217385	.0017762	12.24	0.000	.0182339	.0252431
reg4	1.294163	.462044	2.80	0.006	.3824774	2.205848
urb_reg4	-.0133405	.0065573	-2.03	0.043	-.0262791	-.0004019
_cons	-.4682452	.1007257	-4.65	0.000	-.6669929	-.2694975

التأثير التفاعلي ذو معنوية إحصائية ($p = 0.043$) يشير إلى أن العلاقة بين نسبة السكان الذين يعيشون بالمناطق الحضرية *urban* ولوغاريتم انبعاث ثاني أكسيد الكربون *logco2* تختلف في الدول الأوروبية عنها في بقية دول العالم، التأثير الرئيس للمتغير *urban* هو تأثير إيجابي (0.0217) ويعني أن *logco2* يميل ليكون أعلى في الدول التي يكثر فيها الناس الذين يعيشون في المناطق الحضرية، ولكن التأثير التفاعلي سلبي وهذا يعني أن الميل الأعلى يكون أقل حدة لدول أوروبا، وعليه يمكننا صياغة النموذج أعلاه في معادلتين اثنتين هما:

أوروبا , $reg4 = 1$:

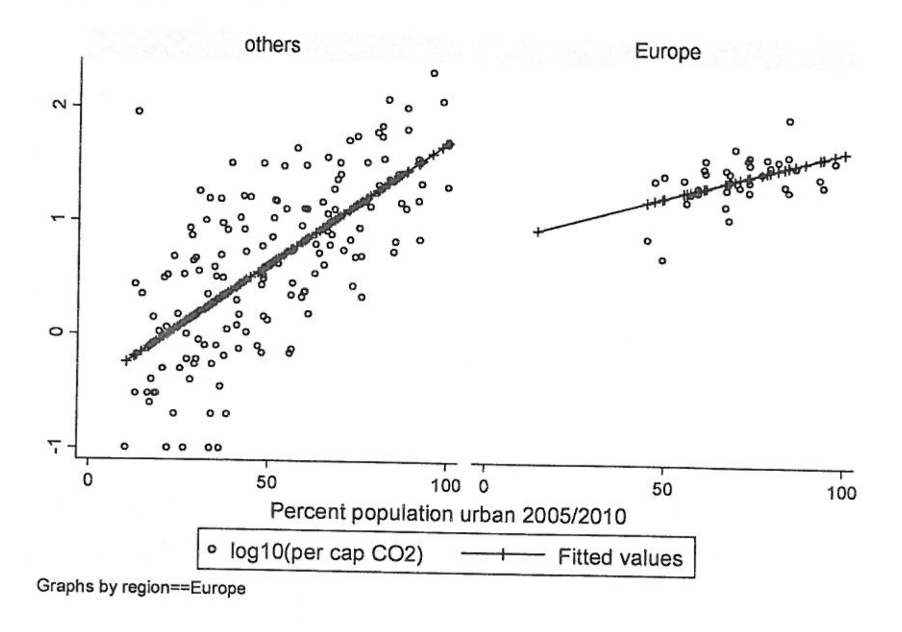
$$\text{المتوقع } logco2 = -0.4682 + 0.0217urban + 1.2942(1) - 0.0133urban(1)$$
$$= -0.4682 + 1.2942 + (0.0217 - 0.0133)urban$$
$$= 0.826 + 0.0084urban$$

الدول الأخرى , $reg4 = 0$:

$$\text{المتوقع } logco2 = -0.4682 + 0.0217urban + 1.2942(0) - 0.0133urban(0)$$
$$= -0.4682 + 0.0217urban$$

بعد إجراء تحليل الانحدار، فإن الأمر **predict** ***newvar*** يقوم بإنشاء متغير جديد يتضمن القيم المتوقعة من آخر انحدار، ويمكن إنشاء رسم بياني للقيم المتوقعة في هذا المثال لإظهار التأثير التفاعلي بيانياً (شكل 7.7)، الخط في الجانب الأيسر ($reg4 = 0$) من الرسم له ميل 0.0217 وتقاطعه مع المحور العمودي y ‎-0.4682، أما الخط في الجانب الأيمن من الرسم ($reg4 = 1$) له ميول أقل حدة (0.0084) ونقطة تقاطع أعلى مع المحور العمودي y (0.826)، النتائج تشير إلى أنه لا توجد دولة أوروبية شهدت انخفاضاً في عدد السكان في المناطق الحضرية، وانخفاضاً في انبعاث ثاني أكسيد الكربون.

```
.predict co2hat
.graph twoway scatter logco2 urban, msymbol(oh)
|| connect co2hat urban, msymbol(+)
|| , by(reg4)
```



الشكل (7.7)

المتغيرات **i.***varname*, **c.***varname* حيث إن الحرف الأول من اسم متغير له إشارة معينة، فالحرف **i** يعني مؤشر (indicator) والحرف **c** يشير إلى

الاستمرارية (continuous) والتي سبق الإشارة إليها في الفصل (6). هذه الحروف تعطي طريقة بديلة لتضمين التفاعلات، الرمز # يحدد التفاعل بين متغيرين اثنين، و## التفاعل المشترك والذي يتضمن تلقائياً كل تفاعلات المستوى الأقل متضمناً هذه المتغيرات، المتغير *reg4* هو عبارة عن متغير مؤشر والمتغير *urban* هو متغير مستمر، إذن نفس النموذج الذي سبق صياغته سابقاً يمكن الحصول عليه بواسطة الأمر:

```
.regress logco2 c.urban i.reg4 c.urban#i.reg4
```

وهذا يكافئ التفاعل العاملي:

```
.regress logco2 c.urban##i.reg4
```

Source	SS	df	MS
Model	55.8882644	3	18.6294215
Residual	46.2772694	181	.255675521
Total	102.165534	184	.555247466

Number of obs =	185
F(3, 181) =	72.86
Prob > F =	0.0000
R-squared =	0.5470
Adj R-squared =	0.5395
Root MSE =	.50564

logco2	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
urban	.0217385	.0017762	12.24	0.000	.0182339	.0252431
1.reg4	1.294163	.462044	2.80	0.006	.3824774	2.205848
reg4#c.urban						
1	-.0133405	.0065573	-2.03	0.043	-.0262791	-.0004019
_cons	-.4682452	.1007257	-4.65	0.000	-.6669929	-.2694975

الأمر **margins** يفهم بأن # أو ## علاقات تفاعلية، نسبة الذين يعيشون في المناطق الحضرية في البيانات تتراوح بين 10% إلى 100%، ويمكننا إنشاء رسم بياني للعلاقات الفاعلية، أولاً نحسب المتوسطات المتوقعة للمتغير *logco2* عند عدة مستويات من المتغير *urban* (10، 40، 70 أو 100) والمتغير *reg4* (0 أو 1).

```
.margins, at(urban = (10(30)100) reg4 = (0 1))
 vsquish
```

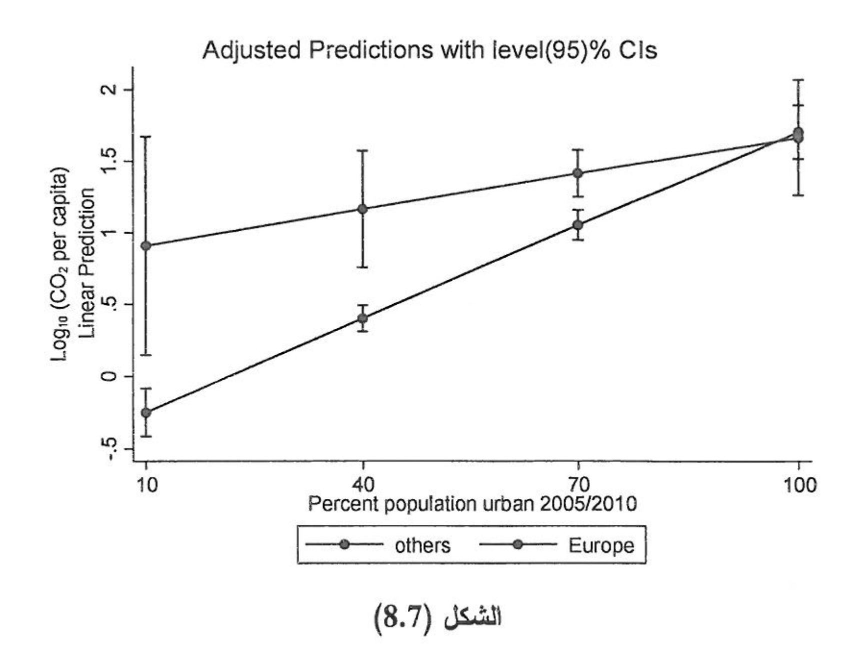
```
Adjusted predictions                                 Number of obs   =        185
Model VCE      : OLS

Expression     : Linear prediction, predict()
1._at          : urban          =          10
                 reg4           =           0
2._at          : urban          =          10
                 reg4           =           1
3._at          : urban          =          40
                 reg4           =           0
4._at          : urban          =          40
                 reg4           =           1
5._at          : urban          =          70
                 reg4           =           0
6._at          : urban          =          70
                 reg4           =           1
7._at          : urban          =         100
                 reg4           =           0
8._at          : urban          =         100
                 reg4           =           1
```

	Margin	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
_at						
1	-.2508599	.084864	-2.96	0.003	-.4171903	-.0845296
2	.9098981	.3890654	2.34	0.019	.147344	1.672452
3	.4012958	.046435	8.64	0.000	.3102849	.4923067
4	1.161839	.2080449	5.58	0.000	.7540788	1.5696
5	1.053452	.052811	19.95	0.000	.9499438	1.156959
6	1.413781	.0831383	17.01	0.000	1.250833	1.576729
7	1.705607	.0953954	17.88	0.000	1.518636	1.892579
8	1.665722	.2055716	8.10	0.000	1.262809	2.068635

الخطوة التالية هي استخدام الأمر **marginsplot** لرسم هذه المتوسطات، لاحظ أنه تم استخدام خيارات مع الأمر **twoway** لتوصيف تفاصيل الرسم البياني، الشكل (8.7) يعرض نفس النموذج الذي يعرضه الشكل (7.7) ولكن بتنسيق مختلف يعرض فترات الثقة للمتوسطات المتوقعة بدلاً من نقاط البيانات، الخيار في الأمر **marginsplot** يحدد l2 (حرف "إل" ثم رقم 2) ثم عنوان المحور في الجانب الأيسر من الرسم.

```
.marginsplot, l2("Log{subscript:10}
 (CO{subscript:2}
per capita)") xlabel(10(30)100)
```



الشكل (8.7)

التأثيرات التفاعلية يمكن أن تتضمن قياسات متغيرين، وهناك طريقة تسمى التمركز تساعد في تقليص مشاكل الارتباط المتعدد multicollinearity مع مثل هذه التفاعلات، وتجعل تأثيراتها الرئيسة أسهل للتفسير. التمركز يتضمن طرح متوسطاتها من المتغيرات قبل تحديد شرط التفاعل كمنتج لهذا التمركز. ومتغيرات التمركز لها متوسطات تساوي صفرًا تقريباً وتكون سالبة للقيم الأقل من المتوسط. الأمر أدناه يحسب إصدار التمركز للمتغيرات *urban, loggdp* ويعطيها اسمه *urban0, loggdp0* وشرط التفاعل *urb_gdp* يتم تعريفه على أنه ناتج *urban0* ضرب *loggdp0*

```
.summarize urban loggdp
```

Variable	Obs	Mean	Std. Dev.	Min	Max
urban	194	55.43488	23.4391	10.25	100
loggdp	179	3.775729	.5632902	2.446848	4.874516

```
.generate urban0 = urban - 55.4
.label variable urban0 "Percent urban,
 centered"
.generate loggdp0 = loggdp - 3.78
```

```
.label variable loggdp0 "log10(GDP per cap), centered"
.generate urb_gdp = urban0 * loggdp0
.label variable urb_gdp "interaction urban0*loggdp0"
```

تمركز أكثر دقة يمكن القيام به باستخدام المتوسطات التي يستخرجها الأمر **summarize**.

```
.summarize urban
.generate urban00 = urban - r(mean)
```

قد نقوم باستبعاد كل المشاهدات التي تحتوي على قيم مفقودة في أي متغيرات عند حساب الانحدار قبل الحصول على المتوسط لغرض التمركز.

حساب انحدار *logco2* على التأثيرات الأساسية المركزية للمتغير *loggdp0* والمتغير *urban0* مع شرط التفاعل *urb_gdp* سوف يوضح أن التأثير التفاعلي سالب وذو معنوية إحصائية.

```
.regress logco2 loggdp0 urban0 urb_gdp
```

Source	SS	df	MS
Model	83.4990753	3	27.8330251
Residual	12.806051	171	.074889187
Total	96.3051263	174	.553477737

Number of obs =	175
F(3, 171) =	371.66
Prob > F =	0.0000
R-squared =	0.8670
Adj R-squared =	0.8647
Root MSE =	.27366

logco2	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
loggdp0	1.116759	.0558107	20.01	0.000	1.006592	1.226925
urban0	.0024787	.0013689	1.81	0.072	-.0002235	.0051809
urb_gdp	-.0082808	.0017418	-4.75	0.000	-.011719	-.0048425
_cons	.8950411	.0267376	33.47	0.000	.8422628	.9478194

نفس الانحدار يمكن حسابه ورسمه للمتغيرات المستمرة باستخدام الأوامر الثلاثة أدناه (النتائج لن يتم عرضها).

```
.regress logco2 c.loggdp0 c.urban0 c.loggdp0#c.urban0
.margins, at(loggdp0 = (-1.3 1.1) urban0 =( -45 45))
.marginsplot
```

التأثيرات الأساسية للانحدار من هذا النوع عند إجراء تمركز للمتغيرات التفاعلية يمكن تفسيره كتأثير لكل متغير عندما تكون قيم المتغيرات الأخرى عند المتوسط، لذا فإن *logco2* المتوقع يزداد بمقدار 1.12 عند زيادة وحدة واحدة في المتغير *loggdp* عندما يكون المتغير *urban* عند قيمته المتوسطة، وبالمثل فإن *logco2* المتوقع يزداد بمقدار بسيط (0.0025) عند زيادة وحدة واحدة في المتغير *urban* عندما يكون المتغير *loggdp* عند قيمته المتوسطة؛ معامل شرط التفاعل *urb_gdp* يوضح بأن كل زيادة وحدة واحدة في عدد السكان الذين يعيشون في المناطق الحضرية تؤدي ضعف تأثير *loggdp* على *logco2* بمقدار 0.008−، كما أن زيادة انبعاث ثاني أكسيد الكربون تزداد بزيادة الثروة، ولكن هذه الزيادة هي أقل حدة في المناطق الأكثر تحضراً.

التقديرات الموثوقة للتباين : Robust Estimates of Variance

الأخطاء المعيارية، واختبارات الفرضيات التي تصاحب الانحدار العادي (مثل الانحدار والتباين) تفترض بأن الأخطاء تتبع التوزيعات المستقلة والمتطابقة، وإذا كان هذا الافتراض غير صحيح، فإنه من المحتمل أن الأخطاء المعيارية سوف تقلل من أهمية التباين من عينة لأخرى، وتعطي فترات ثقة صغيرة وغير واقعية أو احتمالات اختبار منخفضة جداً، للتعامل مع هذه المشكلة والتي تسمى اختلاف التباين heteroskedasticity فإن الأمر **regress** وبعض الأوامر المناسبة الأخرى بها خيار يقوم بتقدير الأخطاء المعيارية بدون الاعتماد على فرضيات قوية أو أحياناً ضعيفة للاستقلالية أو الاعتماد على أخطاء التوزيع المتماثلة، هذا الخيار يستخدم مدخلاً تم اشتقاقه بشكل مستقل من هوبر ووايت وآخرين Huber, White and others ويُشار إليه أحياناً شطيرة مقدّر التباين.

للحصول على معلومات أكثر عن هذا الخيار، قم بطباعة الأمر **help vce option** أو انظر vce_option في دليل المستخدم *Stata Reference Manual* للحصول على تفاصيل تقنية أكثر.

الجزء السابق شرح انحدار *logco2* على ثلاثة متغيرات تنبؤية *loggdp()*، *urban()* ونتائجها *urb_gdp*، ولتكرار نفس الانحدار ولكن مع أخطاء معيارية موثوقة سوف نقوم بإضافة الخيار **vce(robust)**.

```
.regress logco2 loggdp0 urban0 urb_gdp,
 vce(robust)
```

```
Linear regression                               Number of obs =      175
                                                F(  3,    171) =   410.66
                                                Prob > F      =   0.0000
                                                R-squared     =   0.8670
                                                Root MSE      =   .27366
```

logco2	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
loggdp0	1.116759	.0525277	21.26	0.000	1.013072	1.220445
urban0	.0024787	.0013123	1.89	0.061	-.0001116	.005069
urb_gdp	-.0082808	.0016976	-4.88	0.000	-.0116317	-.0049298
_cons	.8950411	.0271616	32.95	0.000	.8414258	.9486564

الجوانب التوصيفية للانحدار – المُعَامِلات وR^2 – تتطابق مع أو بدون الأخطاء المعيارية الموثوقة. ومن ناحية أخرى، فإن الأخطاء المعيارية الموثوقة نفسها مع فترات الثقة واختبارات t وF تختلف عن نظيراتها غير الموثوقة التي رأيناها سابقاً. وعموماً فإن الاختلافات هنا هي اختلافات طفيفة، النتائج الرئيسة في هذا المثال لا تعتمد على افتراض أن الأخطاء مستقلة، وذات توزيع متطابق لكل القيم للمتغيرات التنبؤية.

المنطق الذي تقوم عليه تقديرات الأخطاء المعيارية تم توضيحها في دليل المستخدم لبرنامج ستاتا *User's Gude*، وباختصار نترك الهدف التقليدي وهو تقدير المعالم الصحيحة للمجتمع (β's) لنموذج ما مثل:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

وبدلاً من ذلك، سوف نقوم بمحاولة تحقيق هدف متواضع وهو تقدير التباين من عينة إلى عينة التي قد يوجد في مُعَامِلات b، فإذا قمنا بسحب عينات عشوائية وقمنا بتطبيق OLS بشكل متكرر لحساب قيم b لنموذج مثل:

$$y_i = b_0 + b_1 x_i + e_i$$

نحن لا نفترض أن تقديرات b سوف تقترب من معلمة المجتمع الصحيحة، وفترات الثقة تم إنشاؤها باستخدام الأخطاء المعيارية الموثوقة، وبالتالي نفتقد التفسير التقليدي عند الحصول على احتمال مؤكد (خلال المعاينة المتكررة) التي تحتوي على قيم β الصحيحة، بدلاً من ذلك، فإن فترات الثقة الموثوقة لها احتمال مؤكد (خلال المعاينة المتكررة) تحتوي على قيمة b، والتي تُعرّف بأنها القيمة التي بناءً عليها تقوم العينة b بتقدير التقارب. ولذا فإننا نقلل من الاعتماد على فرضية أخطاء التوزيع المتطابقة من خلال قبول النتائج العادية.

الخيار الآخر للتباين الموثوق هو **vce(cluster *clustervar*)** الذي يسمح لنا بتقليص فرضية الأخطاء المستقلة إلى مستوى محدد وذلك عندما تكون الأخطاء مرتبطة مع مجموعات فرعية أو عنقودية من البيانات. فمثلاً في البيانات المقطعية رأينا اختلافات جوهرية في التباين بين الدول *region* في المثال السابق، إضافة الخيار **vce(cluster *region*)** تؤدي إلى الحصول على الأخطاء المعيارية الموثوقة في المجموعات الفرعية للدول *region*.

```
.regress logco2 loggdp0 urban0 urb_gdp,
 vce(cluster region)
```

```
Linear regression                                      Number of obs =     175
                                                       F(  3,     4) =  771.01
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.8670
                                                       Root MSE      =  .27366

                                   (Std. Err. adjusted for 5 clusters in region)
```

logco2	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
loggdp0	1.116759	.0744462	15.00	0.000	.9100631	1.323454
urban0	.0024787	.0015223	1.63	0.179	-.0017478	.0067052
urb_gdp	-.0082808	.0022161	-3.74	0.020	-.0144336	-.0021279
_cons	.8950411	.0726082	12.33	0.000	.6934485	1.096634

مرة أخرى مُعَامِلات الانحدار وR^2 متطابقة مع تلك التي كانت في النماذج السابقة، ولكن الأخطاء المعيارية وفترات الثقة واختبارات الفرضيات

تغيرت، الأخطاء المعيارية العنقودية أكبر بشكل ملحوظ من تلك التي كانت في النماذج السابقة، وهذا أدى إلى الحصول على إحصائيات *t* أصغر واحتمالات أعلى. استخدام الخيار **vce(robust)** سابقاً أدى إلى تغيرات طفيفة مشيراً إلى عدم وجود مشكلة معينة ومفترضاً أن الأخطاء مستقلة ومتطابقة التوزيع في المتغيرات التنبؤية بالنموذج، استخدام الخيار **vce(cluster *region*)** أدى إلى تغيرات كبيرة، مشيراً إلى أن الأخطاء ليست مستقلة أو متطابقة التوزيع في المتغير *region* وهذا ما كنا نعتقده. وبالتالي فإن تقديرات الخيار **vce(cluster *region*)** معقولة، ويمكن الإفصاح عنها في مكان التقديرات الافتراضية إذا كنا نريد كتابة هذه النتائج في أي بحث.

القيم المتوقعة والبواقي : Predicted Values and Residuals

بعد حساب أي انحدار، فإن الأمر **predict** يمكنه الحصول ليس على القيم المتوقعة فقط، وإنما أيضاً على البواقي وإحصائيات حالات ما بعد التقدير، وهي الإحصائيات التي لها قيم منفصلة لكل مشاهدة في البيانات. في هذا الجزء، سوف ننتقل إلى مثال آخر حول الانحدار البسيط لجليد البحر في المناطق القطبية خلال شهر سبتمبر *area* على السنة *year* (باستخدام ملف البيانات *Arctic9.dta*). الانخفاض يميل ليكون في المتوسط 0.076− أو تقريباً 76000 كم^2 في السنة، وهذا يشرح نحو 75% من التباين في المنطقة القطبية *area* خلال الفترة من 1979 إلى 2011.

```
.use C:\data\Arctic9.dta, clear
.describe year area tempN
```

variable name	storage type	display format	value label	variable label
year	int	%ty		Year
area	float	%9.0g		Sea ice area, million km^2
tempN	float	%9.0g		Annual air temp anomaly 64N-90N C

```
.regress area year
```

Source	SS	df	MS
Model	17.4995305	1	17.4995305
Residual	5.4491664	31	.175779561
Total	22.9486969	32	.717146777

```
Number of obs =      33
F(  1,    31) =   99.55
Prob > F      =  0.0000
R-squared     =  0.7626
Adj R-squared =  0.7549
Root MSE      =  .41926
```

area	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
year	-.0764773	.0076648	-9.98	0.000	-.0921098	-.0608447
_cons	157.4225	15.29154	10.29	0.000	126.2352	188.6098

يمكننا إنشاء متغير جديد باسم *areahat* يحتوي على القيم المتوقعة من هذا الانحدار، ومتغير آخر باسم *areares* يحتوي على بواقي ما بعد التقدير، وذلك باستخدام الأمر **predict**. القيم المتوقعة لها نفس المتوسط الحسابي للمتغير الأصلي *y*، والبواقي لها متوسط صفر $\approx -1.38\times10^{-9}$ = -1.38e09)(0، لاحظ بأن استخدام علامة النجمة في الأمر **summarize** أدناه يعني استخدام كل المتغيرات التي تبدأ باسم "area".

```
.predict areahat
.label variable areahat "Area predicted from year"
.predict areares, resid
.label variable areares "Residuals, area predicted from year"
.summarize area*
```

Variable	Obs	Mean	Std. Dev.	Min	Max
area	33	4.850303	.8468452	3.09	6.02
areahat	33	4.850303	.7395001	3.626667	6.073939
areares	33	-1.38e-09	.4126578	-.8425758	1.116174

القيم المتوقعة والبواقي يمكن تحليلها مثل أي متغيرات أخرى، فمثلاً يمكننا إجراء اختبار الاعتدال للبواقي للتأكد من فرضية الأخطاء الطبيعية، وفي هذا المثال اختبار الالتواء (sktest) يوضح أن البواقي لا تختلف جوهرياً عن التوزيع الطبيعي ($p = 0.45$)

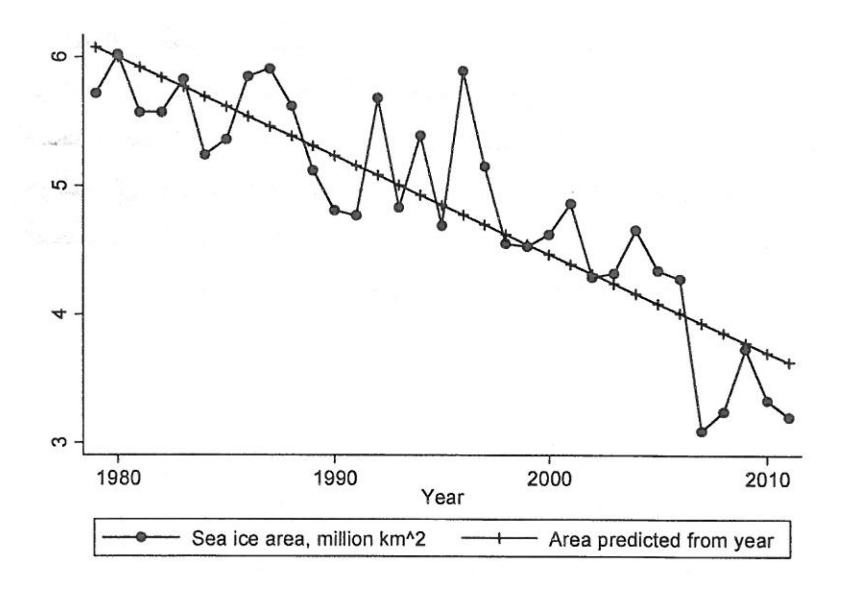
```
.sktest areares
```

Skewness/Kurtosis tests for Normality

Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	joint adj chi2(2)	joint Prob>chi2
areares	33	0.2951	0.5344	1.59	0.4520

إنشاء رسم بياني للقيم المتوقعة مع السنة *year* يوضح خــط الانحــدار (الشكل 9.7).

```
.graph twoway connect area areahat year,
 msymbol(O +)
```



الشكل (9.7)

البواقي تتضمن معلومات عن المناطق التي يكون فيها النموذج ضــعيفاً وهذا يساعد في تشخيص وحل مشاكل التحليل. مثل هذا التحليــل قــد يبــدأ بترتيب واختبار البواقي. البواقي السالبة تظهر عندما يقوم النموذج بإعطـــاء قيم أكبر للقيم المتوقعة، وهذا يحدث في سنوات معينة، حيث تكون المنـــاطق الجليدية أقل من الاتجاه العام للتوقعات. ولبيان السنوات التي بها أقــل مــن خمسة بواقي نقوم بطباعة الأمر:

```
.sort areares
.list year area areahat areares in 1/5
```

	year	area	areahat	areares
1.	2007	3.09	3.932576	-.8425758
2.	2008	3.24	3.856098	-.6160985
3.	1984	5.24	5.691553	-.4515528
4.	2011	3.2	3.626667	-.4266666
5.	1990	4.81	5.232689	-.4226894

ثلاثة من أقل خمسة بواقيٍ حدثت في أحدث خمس سنوات، وهذا يشير إلى التقديرات المبالغ فيها حديثاً.

البواقي الموجبة تظهر عندما تكون القيم الفعلية أعلى من القيم المتوقعة. وحيث إن البيانات تم ترتيبها بواسطة *e* ولعرض أعلى خمسة بواقي نقوم بإضافة محدد "5-" **in -5/l**. هذا المحدد يعني خامس رقم من آخر مشاهدة، وحرف "إل" (لاحظ أنه ليس رقم 1 وإنما حرف إل) يعني آخر مشاهدات، والمحددات **in -5/-1, in 47/1, in 47/51** كل منها يمكنه القيام بنفس الوظيفة.

```
.list year area areahat areares in -5/1
```

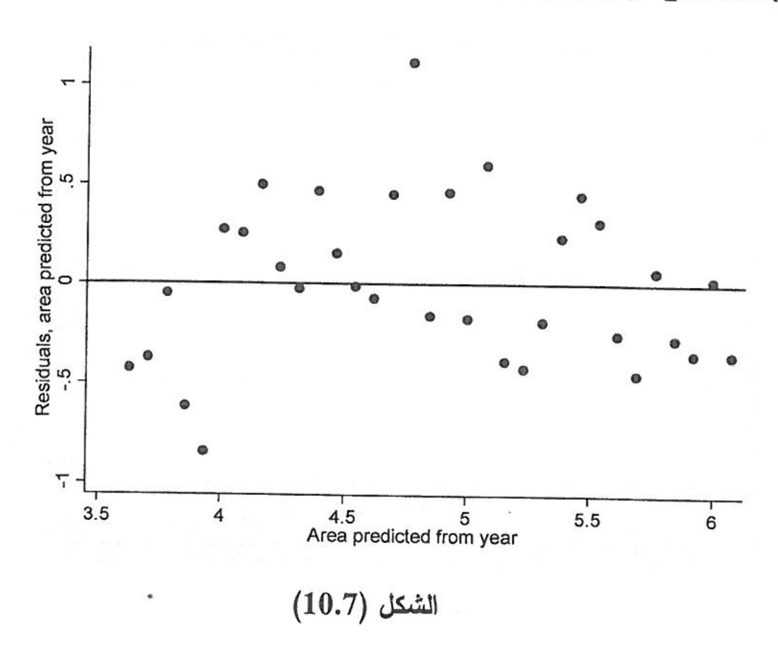
	year	area	areahat	areares
29.	1994	5.39	4.92678	.4632196
30.	2001	4.86	4.391439	.4685608
31.	2004	4.66	4.162007	.4979923
32.	1992	5.68	5.079735	.600265
33.	1996	5.89	4.773826	1.116174

مرة أخرى، هناك نمط معين، حيث إن أعلى بواقي موجبة أو السنوات التي يكون فيها النموذج الخطي أقل جليداً من الجليد الذي تم مشاهدته تمت خلال فترة التسعينيات وحتى بدايات 2000، وقبل الانتقال إلى تحليلات أخرى، يُفترض أن نعيد ترتيب البيانات، وذلك عن طريق الأمر sort *year* وبذلك تكون البيانات مرتبة حسب التسلسل الزمني.

الرسومات البيانية للبواقي مع القيم المتوقعة – والتي تسمى عادة الباقي مقابل المتناسب – تعتبر أداة تشخيصية مناسبة. الشكل (10.7) يوضح مثل هذه الرسومات، حيث يوضح الشكل المتغير *areares* مع المتغير *areahat* مع

خط أفقي تم رسمه عند نقطة الصفر ومتوسط البواقي، لاحقاً في هذا الفصل في الشكل (17.7) يعرض طريقة أخرى لرسم مثل هذه الأشكال البيانية.

```
.graph twoway scatter areares areahat, yline(0)
```



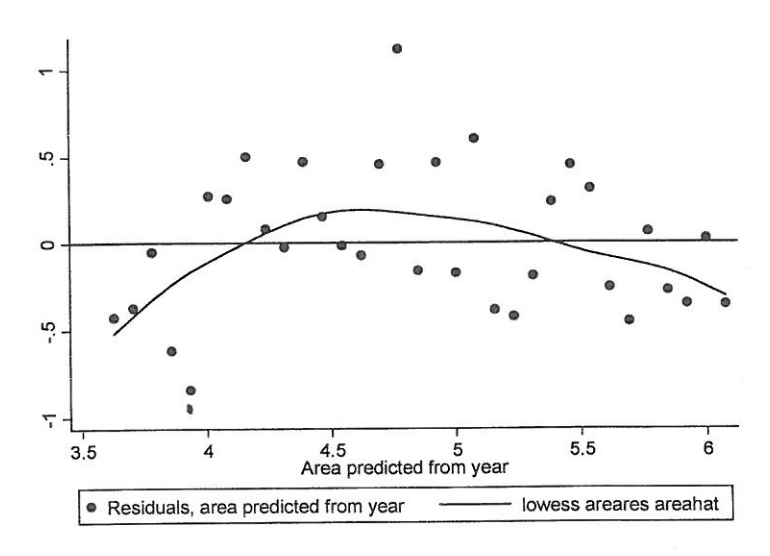
الشكل (10.7)

الأشكال البيانية التي ترسم البواقي مع القيم المتوقعة لاتوضح أي ميول إطلاقاً، فهي مثل سرب من النحل الذي يكون أكثر كثافة في المنتصف (انظر دراسة Hamilton a 1992 للحصول على أمثلة أكثر). ولكن هناك ميولاً مرائياً واضح في الشكل (10.7) حيث إن التوقعات مرتفعة جداً في السنوات الأولى، مما أدى إلى الحصول على بواقي أغلبها سالبة ومنخفضة في المنتصف، ثم بواقي موجبة تميل لتكون مرتفعة جداً في نهاية المدة، حيث تعود البواقي سالبة مرة أخرى، وهذا هو النمط الذي تمت ملاحظته سابقاً عند ترتيب البواقي حسب حجمها.

نمط الارتفاع والانخفاض في البواقي أصبح واضحاً في الشكل (11.7)، حيث إن البواقي مع القيم المتوقعة في الرسم البياني تم وضعهما مع بعضهما بواسطة منحنى منخفض. الانحدار يعتبر طريقة مفيدة لاكتشاف نمط التذبذب

في البيانات. تم الحديث باختصار عن ذلك في الفصل "3" (الـشكل 26.3)، وسوف يتم شرحه بتفصيل أكثر في الفصل (8).

```
.graph twoway scatter areares areahat
 || lowess areares areahat || , yline(0)
```



الشكل (11.7)

هذا النمط غير الخطي في البواقي، يشير إلى أن النموذج الخطي لم يكن مناسباً لهذه البيانات – سوف نعود لهذه النقطة في الجزء القادم – كما أننا سنعود إليها مرة أخرى في الفصل (8).

بعد أي تحليل انحدار، فإن برنامج ستاتا يحفظ المُعامِلات وتفاصيل أخرى بشكل مؤقت، ولذا فإن _b[*varname*] يشير إلى معامل متغير مستقل *varname* بينما _b[_cons] يشير إلى معامل _cons (عادة التقاطع مع المحور العمودي *y*).

```
.display _b[year]
  -.07647727
.display _b[_cons]
 157.42247
```

بالرغم من أن الأمر **predict** يسهّل عملية احتساب القيم المتوقعة والبواقي، فإنه من الممكن تعريف نفس المتغيرات من خلال زوج من أوامر

generate باستخدام معاملات [ ]_b، والمتغيرات الناتجة والتي تم تسميتها *areahat1* و *areres1* أدناه لها نفس خصائص القيم المتوقعة والبواقي الناتجة من الأمر predict ولكن لأغراض معينة، فإن الأمر generate يعطي المستخدمين مرونة أكثر.

```
.generate areahat1 = _b[_cons] + _b[year]*year
.generate reares1 = area - areahat1
.summarize area*
```

Variable	Obs	Mean	Std. Dev.	Min	Max
area	33	4.850303	.8468452	3.09	6.02
areahat	33	4.850303	.7395001	3.626667	6.073939
areares	33	-1.38e-09	.4126578	-.8425758	1.116174
areahat1	33	4.850303	.7395001	3.626667	6.073939

حالات إحصائية أخرى : Other Case Statistics

الأمر predict يمكنه حساب العديد من الإحصائيات الأخرى المناسبة للنماذج، وبعد الأمر regress (أو anova) الخيار predict يتضمن التالي (سوف يتم التعويض عن أي اسم متغير جديد بـ *new* في هذه الأمثلة).

`predict new`	القيم المتوقعة للمتغير *y*، predict *new*, xb يعني نفس الشيء (الإشارة إلى xb، قوة قيمة *y* المتوقعة).
`predict new, resid`	البواقي.
`predict new, rstandard`	البواقي القياسية.
`predict new, rstudent`	البواقي القياسية التي تقيس *i* تأثير المشاهدة على تقاطع المحور العمودي *y*
`predict new, stdp`	الأخطاء المعيارية للمتوسط المتوقع *y*.
`predict new, stdf`	الأخطاء المعيارية لكل قيمة فردية متوقعة *y*، بعض الأحيان تسمى الأخطاء المعيارية للتقدير أو الأخطاء المعيارية للتوقع.
`predict new, hat`	العناصر المحورية لمصفوفة التقدير (كما أن الأمر leverage يقوم بنفس الوظيفة).

`predict new, cooksd`	تأثير مسافة كوك *D* Cook's والذي يُقاس بتأثير المشاهدة *i* في كل المعاملات التي في النموذج (أو بشكل مكافئ كل المشاهدات *n* للقيم المتوقعة *y*).

هناك خيارات أكثر يمكنها الحصول على الاحتمالات المتوقعة والقيم المتوقعة. وللحصول على قائمة بهذه الخيارات قم بطباعة الأمر **help regress**.

لشرح هذه الخيارات، سوف نعود لبيانات الجليد بالقطب الشمالي بملف البيانات *Arctic9.dta*. تحليل البواقي في الجزء السابق أشار إلى نموذج خطي في الشكل (9.7) وهو ليس مناسباً لهذه البيانات. أحد البدائل البسيطة الأخرى، والذي يوصف بأنه الانحدار من الدرجة الثانية، وهو يتضمن تحليل انحدار المتغير التابع على *year* ومربع *year*، ويمكننا أن نبدأ ذلك بإنشاء متغير جديد يساوي مربع *year*.

```
.generate year2 = year^2
.regress area year year2
```

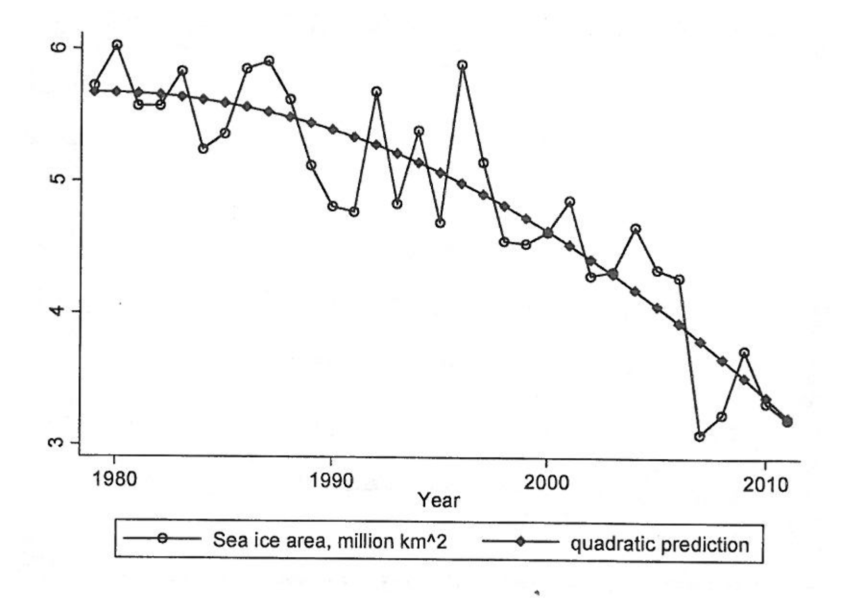
```
      Source |       SS       df       MS              Number of obs =      33
-------------+------------------------------           F(  2,    30) =   67.73
       Model |  18.7878137     2  9.39390686           Prob > F      =  0.0000
    Residual |  4.16088316    30  .138696105           R-squared     =  0.8187
-------------+------------------------------           Adj R-squared =  0.8066
       Total |  22.9486969    32  .717146777           Root MSE      =  .37242

------------------------------------------------------------------------------
        area |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |   9.658356   3.194155     3.02   0.005     3.135022    16.18169
       year2 |  -.0024398   .0008005    -3.05   0.005    -.0040747   -.0008049
       _cons |  -9552.853   3186.119    -3.00   0.005    -16059.78   -3045.931
------------------------------------------------------------------------------
```

الانحدار من الدرجة الثانية يعرض تحسناً عاماً في تناسب R^2_a=0.8066 بالمقارنة مع 0.7549 للانحدار الخطي. وعلى كل حال، فإن التحسن هو تحسن ذو معنوية إحصائية، كما يشير إلى ذلك معامل المتغير (*year*(*p*=0.005. الرسم البياني في الشكل (12.7) يعرض هذا التحسن. والنموذج من الدرجة الثانية ليس مرتفعاً بشكل مستمر، حيث إنه أعلى ثم أقل من الجليد المتوقع، كما حدث في النموذج الخطي السابق.

```
.predict areahat2
```

```
.label variable areahat2 "quadratic prediction"
.graph twoway connect area areahat2 year,
 msymbol(Oh d)
```



الشكل (12.7)

هناك بديلان للقيام بنفس الانحدار من الدرجة الثانية، فبدون القيام بإنشاء متغير جديد للجذر التربيعي لمتغير *year* يمكننا استخدام الرمز التفاعلي ببرنامج ستاتا (#). الأوامر الثلاثة أدناه تقوم بتقدير نفس النموذج. سوف نحتاج للرمز # بعد الأمر **margins** للقيام بالمهمة بالطريقة المطلوبة.

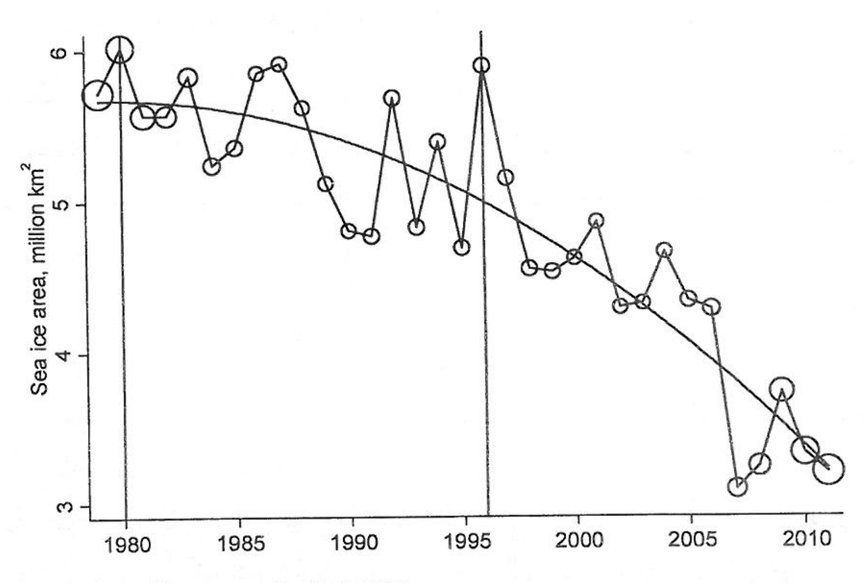
```
.regress area year area2
.regress area year c.year#c.year
.regress area c.year##c.year
```

بالرغم من التحسن الواضح في الشكل (12.7)، فإن الانحدار من الدرجة الثانية يمكنه أن ينتج أو يزيد من نفس المشاكل الإحصائية، ومن الأمثلة على هذه المشاكل: القدرة على التأثير، والذي يعني التأثير المحتمل للمشاهدات التي بها قيم استثنائية، والتي يمكن الحصول عليها باستخدام الأمر **predict**. في هذا المثال، قمنا بتسمية مقياس التأثير باسم *leverage*

```
.predict leverage, hat
```

الشكل (13.7) يعرض منحنى الانحدار من الدرجة الثانية مرة أخرى، في هذه المرة الشكل يجعل رموز الربط تظهر متناسبة مع قوة التأثير، وذلك بتحديد الأوزان التحليلية [aw=*leverage*]، كما أن الشكل يضع منحنى متوسط (**mspline**) يربط القيم المتوقعة *areahat2*. هذا المنحنى في العادة يُعد أفضل من الخط أو الخط المتصل لتوضيح العلاقات في الرسم البياني، لذا فهو يساعد في تحديد رقم أكبر للنطاقات في الرسم البياني. وللحصول على مزيد من المعلومات عن أنواع الرسم البياني **twoway** قم بطباعة الأمر **help twoway mspline**.

```
.graph twoway connect area year [aw=leverage],
 msymbol(Oh)
 || mspline areahat2 year, bands(50)
 lwidth(medthick)
 || ,note("Data points area proportional to
 leverage")
 legend(off) xline(1980 1996)
 xlabel(1980(5)2010, grid)
 xtitle("")
 ytitle("Sea ice area, million
 km{superscript:2}")
```



الشكل (13.7)

الرموز في الشكل (13.7) تكون أكبر حجماً للسنة الأولى والسنة الأخيرة، لأن هذه السنوات لها تأثيرات أعلى، والانحدار من الدرجة الثانية يميل ليؤكد على أهمية قيم x المتطرفة (حتى ولو قمنا بتربيع هذه القيم، فإنها تظل متطرفة) لذا فإن النموذج يحاول تتبع هذه القيم، لاحظ كيف أن المنحنى يتناسب مع السنة الأولى والسنة الأخيرة.

القدرة على التأثير تعكس احتمالية التأثير. والإحصائيات الأخرى تقيس التأثير بشكل مباشر. أحد إحصائيات هذا التأثير هو DFBETAS، والتي تشير بطريقة ما إلى مُعاملات الأخطاء المعيارية للمتغير $x1$ سوف تتغير إذا تم استبعاد المشاهدات i من تحليل الانحدار. يمكن القيام بذلك لمتغير تنبؤي واحد من خلال الأمر **predict** ***new*,dfbeta(*xvarname*)**، قيم DFBETAS لكل المتغيرات التنبؤية في النموذج يمكن الحصول عليها بسهولة عن طريق الأمر **dfbeta** والذي يقوم بإنشاء متغير جديد لكل متغير تنبؤي. وفي مثالنا هناك متغيران تنبؤيان هما *year, year2* وبالتالي فإن الأمر **dfbeta** يعرف إحصائيات مؤثرة جديدة واحدة لكل متغير تنبؤي.

```
.dfbeta
.describe _dfbeta*
```

variable name	storage type	display format	value label	variable label
_dfbeta_1	float	%9.0g		Dfbeta year
_dfbeta_2	float	%9.0g		Dfbeta c.year#c.year

```
. summarize _dfbeta*
```

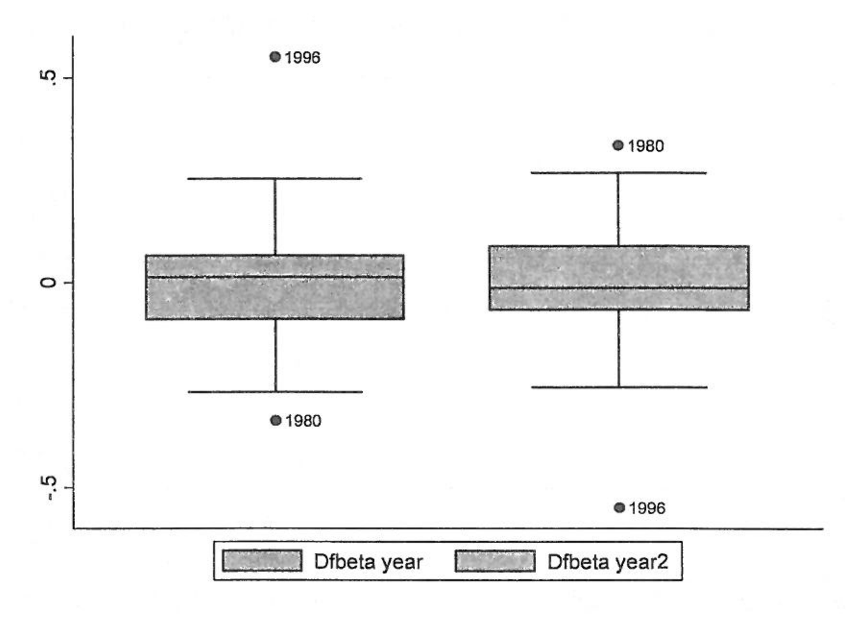
variable name	storage type	display format	value label	variable label
_dfbeta_1	float	%9.0g		Dfbeta year
_dfbeta_2	float	%9.0g		Dfbeta year2

```
. summarize _dfbeta*
```

Variable	Obs	Mean	Std. Dev.	Min	Max
_dfbeta_1	33	.0007181	.1702339	-.3360457	.550404
_dfbeta_2	33	-.0007234	.1702921	-.550294	.3353676

رسم الصندوق في الشكل (14.7) يوضح توزيعات متغيرات DFBETAS مع قيم متطرفة تم تسميتها باسم *year*.

```
.graph box _dfbeta*, marker(1, mlabel(year))
marker(2, mlabel(year))
```



الشكل (14.7)

قيم المتغير *dfbeta_1_* تقوم بقياس تأثير كل مشاهدة على معامل المتغير *year*، سنة 1996 (والتي تعرض قيمة مرتفعة في الشكل 13.7) وهي التي استحوذت على أغلب التأثير في معامل السنة *year*، قيمة المتغير *dfbeta_1=0.55_* 0.55 توضح بأن معامل المتغير *year* في الانحدار للعينة بالكامل *أن حوالي 0.55 من الأخطاء المعيارية أعلى من المفترض في حالة لو ما قمنا بإعادة تقدير النموذج مع استبعاد سنة 1996*. وبالمثل فإن القيم السالبة لسنة 1980 وهي *dfbeta_1=-0.336_* في معامل المتغير *year* تشير إلى أن حوالي 0.336 من الأخطاء المعيارية أقل من المفترض في حالة لو ما قمنا بإعادة تقدير النموذج مع استبعاد سنة 1980، وبينما كان معامل *year* أعلى من المفترض في سنة 1996، إلا أن معامل *year2* أقل بمقدار مشابه،

والعكس قابل للتطبيق في سنة 1980؛ في البيانات الأصلية (الشكل 13.7) لم تكن سنة 1980 متطرفة ولكنها ذات تأثير أعلى مما جعل سنة 1980 أكثر تأثيراً من السنوات الأخرى التي تأثيرها أقل.

إذا لم نكن متأكدين من كيفية قراءة DFBETAS يمكننا تأكيد تفسيرنا لهذه القيم من خلال إعادة تحليل الانحدار مع استبعاد القيم المؤثرة. ولتوفير مساحة، فإن الأمر **quietly** هنا يقوم بالتأكيد على مخرجات الانحدار، ويعرض فقط مُعَاملات المتغير *year* والمتغير *year2* وبينهما مسافات خالية حتى نتمكن من قراءتها بوضوح. تحليل الانحدار الأول يستخدم العينة بالكامل، بينما تحليل الانحدار الثاني يستبعد سنة 1996، والتي كان بها معامل المتغير *year* أقل (من 9.658 إلى 8.067) ومعامل المتغير *year2* أعلى (من 0.0024- إلى 0.0020-)، إن عملية استبعاد سنة 1980 في تحليل الانحدار الثالث لها تأثير معاكس، حيث إن معامل المتغير *year* أصبح أعلى (من 9.658 إلى 10.730) ومعامل المتغير *year2* أصبح أقل (من 0.0024- إلى 0.0027).

```
. quietly regress area year year2
. display _b[year] " " _b[year2]
 9.6583565 -.00243981

.quietly regress area year year2 if year!=1996
.display _b[year] " " _b[year2]
 8.0666424 -.00204096

.quietly regress area year year2 if year!=1980
.display _b[year] " " _b[year2]
 10.730059 -.00270786
```

إذا قمنا باستخدام # لإدخال شرط تربيع، فإن آخر أمر سوف يكون كما يلي:

```
.quietly regress area  year c.year#c.year if
 year!-1980
.display _b[year] " " _b[c.year#c.year]
 10.730059  -.00270786
```

التأثير أو DFBETAS أو أي إحصائيات أخرى يمكن استخدامها مباشرة لاستبعاد مشاهدات من أي تحليل، فمثلاً أول أمرين من الأوامر أدناه تقوم

بحساب مسافة كوك Cook's D وهي إحصائية تقوم بقياس تأثير كل مشاهدة على النموذج ككل بدلاً من المُعامِلات الفردية كما يحدث في DFBETAS، الأمر الثالث من الأوامر أدناه يكرر تحليل الانحدار الأولي باستخدام المشاهدات التي ظهرت في مسافة كوك فقط وهي أقل من 0.10

```
.regress area year year2
.predict D, cooksd
.regress area year year2 if D< .10
```

باستخدام أي تعريف ثابت عن سبب ظهور القيم المتطرفة لأنه من الممكن أن نرى الكثير منها في العينات الأكبر حجماً، لهذا السبب فإن تحديد أحجام العينات أمر مرغوب فيه لتحديد المشاهدات الاستثنائية، وبعد تحديد نموذج الانحدار المناسب مع مُعَامِلات K (هذا يتضمن تحديد المعامل الثابت). بناءً على عدد المشاهدات n يمكننا فحص هذه المشاهدات لمعرفة أي من الشروط أدناه صحيح:

التأثير leverage $h > 2K/n$

مسافة كوك Cook's $D > 4/n$

المُعَامِلات المناسبة $DFITS > 2\sqrt{\frac{K}{n}}$

Welsch's $W > 3\sqrt{K}$

$DFBETA > 2/\sqrt{n}$

$|COVRATIO - 1| \geq 3K/n$

سبب اختيار الحدود، وإحصائيات التشخيص أعلاه، يمكن الحصول عليها من دراسات Belsley, Kuh and Welsch (1980, 1982) Cook and Weisberg(1982, 1994) أو Fox (1991).

تشخيص الارتباط المتعدد واختلاف التباين :

Diagnosing Multicollinearity and Heteroskedasticity

الارتباط المتعدد يشير إلى مشكلة العلاقات الخطية القوية جداً بين المتغيرات التنبؤية أو المتغيرات المستقلة في النموذج. إذا وجدت ارتباطاً متعددًا كاملاً بين المتغيرات التنبؤية، فإن معادلة الانحدار سوف تفتقر إلى

الحلول المطلوبة. برنامج ستاتا يحذرنا، ثم يقوم باستبعاد المتغير التنبؤي الذي يسبب المشكلة. وفي حالة وجود ارتباط متعدد مرتفع وليس كاملاً، فإن هذا يؤدي إلى مشاكل طفيفة جداً. وإذا قمنا بإضافة متغير تنبؤي جديد له علاقة قوية مع المتغيرات التنبؤية الموجودة مسبقاً في النموذج، فإن ذلك سوف يؤدي إلى ظهور أعراض لمشاكل أخرى منها ما يلي:

- أخطاء معيارية مرتفعة بشكل جوهري مع إحصائيات t منخفضة.
- تغير غير متوقع في مقدار المُعامِلات أو إشاراتها.
- مُعَامِلات غير جوهرية بالرغم من وجود R^2 مرتفعة.

الانحدار المتعدد يحاول تقدير التأثيرات المستقلة لكل متغير x، وهناك معلومات قليلة للقيام بذلك إذا كان واحد أو أكثر من المتغيرات x ليس لها تباين مستقل، الأعراض التي تم ذكرها أعلاه هي عبارة عن تحذير بأن تقديرات المُعَامِلات أصبحت غير موثوقة، وربما تغيرت بشكل كبير مع تغييرات طفيفة في البيانات أو النموذج نفسه. هناك حاجة إلى المزيد من استكشاف الأخطاء، ومحاول تشخيصها لتحديد ما إذا كان الارتباط المتعدد مشكلة حقيقية. وإذا كانت كذلك فماذا يجب أن يتم فعله حيالها؟

مصفوفات الارتباط في العادة تعتبر كمؤشر تشخيصي، ولكن لها حدود معينة لاكتشاف الارتباط الخطي المتعدد. الارتباطات يمكنها اكتشاف الارتباط الخطي أو العلاقات الخطية بين زوج من المتغيرات. ومن ناحية أخرى، فإن الارتباط المتعدد يتضمن علاقات خطية بين مجموعة من المتغيرات المستقلة، والتي قد لاتكون واضحة من الارتباطات. التشخيص المناسب يتوفر من خلال الأمر **estat vif**, لتضخم التباين.

```
.regress area year year2

      Source |       SS       df       MS              Number of obs =      33
-------------+------------------------------           F(  2,    30) =   67.73
       Model |  18.7878137     2  9.39390686           Prob > F      =  0.0000
    Residual |  4.16088316    30  .138696105           R-squared     =  0.8187
-------------+------------------------------           Adj R-squared =  0.8066
       Total |  22.9486969    32  .717146777           Root MSE      =  .37242

------------------------------------------------------------------------------
        area |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |   9.658356   3.194155     3.02   0.005     3.135022    16.18169
       year2 |  -.0024398   .0008005    -3.05   0.005    -.0040747   -.0008049
       _cons |  -9552.853   3186.119    -3.00   0.005    -16059.78   -3045.931
------------------------------------------------------------------------------
```

```
.estat vif
```

Variable	VIF	1/VIF
year	220094.55	0.000005
year2	220094.55	0.000005
Mean VIF	220094.55	

العمود 1/VIF في اليمين بالجدول أعلاه، يعطي قيماً تساوي $1 - R^2$ من انحدار كل متغير x على المتغيرات الأخرى x. ففي مثالنا عن الجليد في القطب الشمالي، تمثل حالة شاذة، حيث إن 0.000005 – وهي أقل من 0.0005 في المائة لتباين متغير *year* هي مستقلة عن المتغير المستقل *year2* والعكس. كما أن عامل تضخم التباين أو قيم VIF نفسها مع المتغيرين التنبؤيين في النموذج يوضح أن التباين في مُعامِلاتها أعلى من المفترض بحوالي 220000 مرة، الأخطاء المعيارية للمتغير *year* في النموذج الخطي البسيط الذي رأيناه سابقاً في هذا الفصل تساوي 0.0076648 والخطأ المعياري المقابل في النموذج من الدرجة الثانية أعلاه أعلى بمرات عديدة من 3.194155.

قيم VIF: تشير إلى وجود ارتباط خطي خطير، وحيث إنه يوجد لدينا متغيران تنبؤيان في هذا المثال، فإننا نستطيع تأكيد ذلك من خلال النظر إلى الارتباطات. النتائج أدناه تظهر بأن المتغيرين *year, year2* مرتبطان بشكل كامل.

```
.correlate area year year2
```

(obs=33)

	area	year	year2
area	1.0000		
year	-0.8732	1.0000	
year2	-0.8737	1.0000	1.0000

الارتبط الخطي أو الارتباط الخطي المتعدد، يمكن أن يحدث في أي نوع من النماذج، ولكنهما أكثر انتشاراً في النماذج التي تكون فيها بعض المتغيرات التنبؤية واضحة عن الأخرى، مثل نماذج التأثيرات التفاعلية أو نماذج الانحدار من الدرجة الثانية. الحل البسيط وهو التمركز– الذي تمت

الإشارة إليه سابقاً في شروط التفاعل – يمكنه أيضاً أن يساعد في الانحدار من الدرجة الثانية، لتقليل الارتباط الخطي المتعدد، فإن التمركز في العادة هو الحل، حيث إنه يُنتج معاملات أكثر دقة مع أخطاء معيارية منخفضة.

يمكننا القيام بالتمركز من خلال طرح المتوسط من أحد متغيرات *x* قبل حساب الثاني، متوسط السنة *year* في هذه البيانات هو 1995، والتمركز لهذا المتغير يسمى *year0* ويمثل السنوات قبل 1995 (سالب) أو ما بعد سنة 1995 (موجب)، المتغير *year0* المتمركز له متوسط صفر، والمتغير الثاني الجديد *year02* يساوي مربع *year0* وقيمه تتراوح ما بين 256 (عندما تكون *year0*=-16، وهذا يعني 1979) إلى صفر (عندما *year0* = 0 وهذا يعني 1995) ثم يعود من جديد ليكون 256 (عندما *year0* = +16 وهذا يعني 2011).

```
.gen year0 = year - 1995
.gen year02 = year0 ^2
.summarize year year0 year02
```

Variable	Obs	Mean	Std. Dev.	Min	Max
year	33	1995	9.66954	1979	2011
year0	33	0	9.66954	-16	16
year02	33	90.66667	82.23847	0	256

بعد التمركز سوف تتضح لدينا الفروقات في معامل المتغير المتمركز والخطأ المعياري، ولمشاهدة هذا سوف نقوم بتحليل انحدار جليد البحر *area* على المتغير *year0* والمتغير *year02*. النتائج تشير إلى أن R^2 واختبار F هي بالضبط نفس النتائج التي أظهرها تحليل انحدار *area* على المتغيرات التي لم يتم إجراء تمركز لها وهي *year, year2* كما أن القيم المتوقعة والبواقي استمرت كما هي. أما في نتائج المتغيرات المتمركزة فإن الأخطاء المعيارية للمتغير *year0* أقل بكثير، وفترات الثقة أصبحت أضيق، وإحصائيات t أكبر، وبواقي اختبار t أصبحت ذات معنوية أكبر مما كانت عليه في تحليل الانحدار غير المتمركز مع متغير *year* الأصلية.

```
.regress area year0 year02
```

Source	SS	df	MS
Model	18.7878137	2	9.39390686
Residual	4.16088316	30	.138696105
Total	22.9486969	32	.717146777

```
Number of obs =      33
F(  2,    30) =   67.73
Prob > F      =  0.0000
R-squared     =  0.8187
Adj R-squared =  0.8066
Root MSE      =  .37242
```

area	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
year0	-.0764773	.0068085	-11.23	0.000	-.0903821	-.0625725
year02	-.0024398	.0008005	-3.05	0.005	-.0040747	-.0008049
_cons	5.071512	.0973195	52.11	0.000	4.872759	5.270265

بالرغم من أن المتغير *year* والمتغير *year2* مرتبطان بالكامـل، فـإن ارتباطهما بعد التمركز أصبح أقرب للصفر.

.correlate *year0 year02*

(obs=33)

	year0	year02
year0	1.0000	
year02	-0.0000	1.0000

ولأن المتغيرين التنبؤيين غير مرتبطين ولم يحدث هناك تضخم للتباين.

.estat vif

Variable	VIF	1/VIF
year0	1.00	1.000000
year02	1.00	1.000000
Mean VIF	1.00	

كما أن الأمر **estat** يقوم بحساب إحصائيات تشخيصية أخـرى مفيـدة، فمثلاً **estat hettest** يقوم باختبار اختلاف التباين لافتـراض تبـاين الخطـأ المستمر، حيث يقـوم بذلك من خلال اختبـار ما إذا كانت البواقي المعيارية المربعة ترتبـط ارتباطاً خطياً بالقيـم المتوقعـة (انظر دراسـة Cook and Weisberg1994 لمزيد من الأمثلة والشرح). وكما هو موضح أدناه فإن الأمر

estat hettest لا يعطي سبباً لرفض فرضية العدم التي تقول بثبات التبـــاين، حيث إن النتائج لاتوضح أن اختلاف التباين ذو معنوية إحصائية.

```
.estat hottest

 Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of area

         chi2(1)      =     0.00
         Prob > chi2  =   0.9802
```

ومن ناحية أخـــرى، فإن وجود اختلاف تباين ذو معنوية إحصائية يعني أن الأخطاء المعيارية يمكن أن تكون متحيّزة، وبالتالي فإن نتـــائج اختبـــار الفرضيات تكون غير صالحة.

نطاقات الثقة في الانحدار البسيط :

Confidence Bands in Simple Regression

هذا الجزء يشرح بعض الرسومات البيانية الإضافية التي تـــساعد فـــي العرض المرئي لنماذج الانحدار أو تشخيص المشاكل المحتملة، بالعودة إلى ملف البيانات *Arctic9.dta*، نجد أن المتغير *tempN* يوضح متوسط القيم غيـــر العادية السنوية لدرجة حرارة الهواء للمنطقة الشمالية بالكامل من 64 إلـــى 90 درجة شمالاً (من سجلات وكالة ناسا لدرجات حرارة الهـــواء وســـطح البحر)، القيم غير العادية في درجات الحرارة تمثل الاختلافـــات بالـــدرجات المئوية خلال الفترة من 1951-1980، القيم غير العادية الموجبة هي قـــيم أعلى من متوسط درجات الحرارة خلال الفترة 1951-1980.

سبق لنا وأن رأينا هذه المناطق ومداها وحجمها فـــي جليـــد المنـــاطق الشمالية (خصوصاً في شهر سبتمبر) شهدت انخفاضاً خلال الفتـــرة 1979-2011 وهي فترة المراقبة بالأقمار الصناعية. وليس مفاجئاً بـــأن درجـــات حرارة هواء سطح البحر شهدت ارتفاعاً خلال نفس المدة بالرغم من أن هذا ليس السبب الوحيد في انخفاض الجليد. الميل في ارتفاع درجـــات الحـــرارة

يزداد بمقدار حوالي 0.058 درجة مئوية في السنة أو 0.58 في العقد (عشر سنوات)، وهو معدل أسرع من معدل الارتفاع العالمي ككل. وللمقارنة، فإن بيانات ناسا – لم يتم عرضها هنا – تشير إلى أن الميل في ارتفاع درجات الحرارة كان بمعدل 0.16 درجة مئوية خلال هذه السنوات.

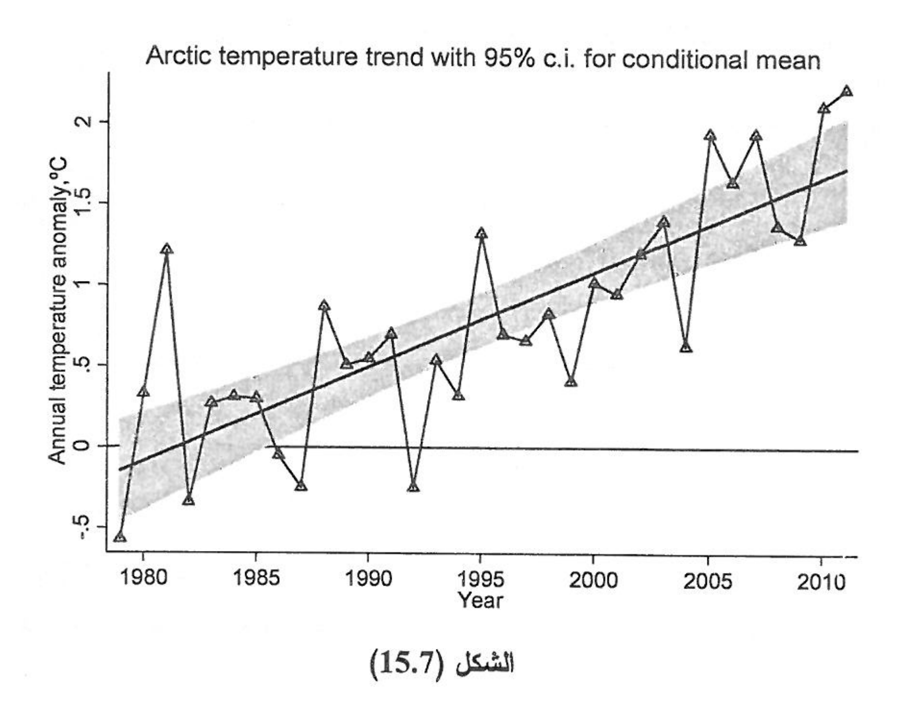
```
.regress tempN year

      Source |       SS       df       MS              Number of obs =      33
-------------+------------------------------           F(  1,    31) =   51.64
       Model |  10.2449886     1  10.2449886           Prob > F      =  0.0000
    Residual |  6.15050844    31  .198403498           R-squared     =  0.6249
-------------+------------------------------           Adj R-squared =  0.6128
       Total |   16.395497    32  .512359282           Root MSE      =  .44543

------------------------------------------------------------------------------
       tempN |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |    .058516   .0081432     7.19   0.000     .0419079    .0751242
       _cons |  -115.9492   16.24582    -7.14   0.000    -149.0828   -82.81563
------------------------------------------------------------------------------
```

الشكل (15.7) يوضح الميل نحو الارتفاع في القيم غير العادية لدرجات الحرارة في المنطقة القطبية الشمالية، درجات الحرارة الواقعية تم تمثيلها في الرسم البياني ووضعها على خط الانحدار مع فترة ثقة 95% للمتوسط الشرطي الذي تم تحديده بواسطة الأمر **twoway lfitci, stdp**. الخيارات الأخرى تقوم بتحديد سُمك خط الانحدار ليكون متوسط السُمك ويكون عنوان الرسم مكتوباً بنص ذي حجم متوسط. كما تم استخدام رمز درجة مئوية وهو الرمز ASCII رقم 186 تم إدراجه في عنوان المحور العمودي (انظر الشكل "16.3" في الفصل "3" لمعرفة الرموز الأخرى في ASCII).

```
.graph twoway lfitci tempN year, stdp
 lwidth(medthick)
 || connect tempN year, msymbol(Th)
 || , ytitle("Annual temperature
 anomaly,`=char(186)'C")
 legend(off) xlabel(1980(5)2010) yline(0)
 title("Arctic temperature trend with 95% c.i.
 for
 conditional mean", size(medlarge))
```



الشكل (15.7)

العديد من القيم السنوية تقع خارج فترات الثقة في الشكل (15.7) مما يؤكد حقيقة أن هذه الفترات تشير إلى قيم المتوسط الشرطية أو الميل نفسه بدلاً من التوقعات الفردية. فلو افترضنا أننا نريد الحصول على توقّع فردي لسنة 2012 وإيجاد فترة الثقة المناسبة لهذا التوقع، فإحدى الطرق للقيام بذلك هي استخدام محرر البيانات Data Editor لإضافة 34 صفاً جديدًا من البيانات. وتحتوي هذه البيانات على قيم سنة 2012، أو يمكن القيام بذلك من خلال الأمرين أدناه:

```
.set obs 34
.replace year = 2012 in 34
```

ثم قم بتكرار الانحدار للحصول على القيم المتوقعة، والأخطاء المعيارية للتوقعات (**stdf**). القيم التي تقع أعلى أو أدنى من حدود الثقة 95% هي القيم المتوقعة تقريباً ناقص أو زائد مرتين الخطأ المعياري للتوقعات: *tempNhat* ناقص أو زائد 2**tempNse*.

```
.predict tempNhat
.label variable tempNhat "Predicted temperature"
```

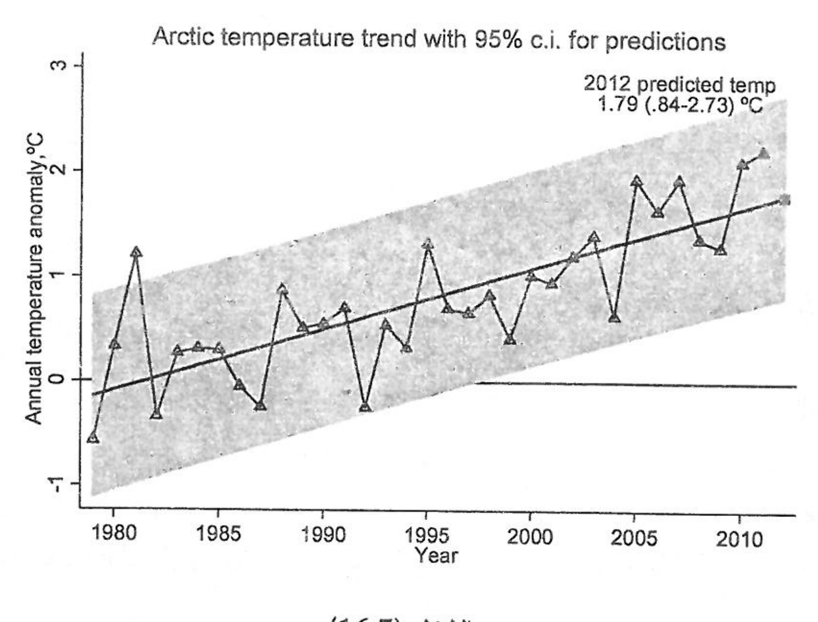
```
.predict tempNse, stdf
.label variable tempNse "Standard error of
 forecast"
.gen tempNlo = tempNhat - 2*tempNse
.label variable tempNlo "lower confidence limit"
.gen tempNhi = tempNhat + 2*tempNse
.label variable tempNhi "upper confidence limit"
.list year tempN* in -5/1
```

	year	tempN	tempNhat	tempNse	tempNlo	tempNhi
30.	2008	1.37	1.551012	.4643515	.6223085	2.479715
31.	2009	1.29	1.609528	.4662754	.6769768	2.542078
32.	2010	2.11	1.668044	.468333	.7313777	2.60471
33.	2011	2.22	1.72656	.4705225	.7855148	2.667605
34.	2012	.	1.785076	.4728422	.8393915	2.73076

يمكننا الآن إنشاء رسم بياني لقيم المتغيرات *tempNlo* و*tempNhi* في منطقة المدى (**twoway rarea**) ورؤوس مدببة (**rspike**) ورؤوس مدببة مغطاة (**rcap**) أو رسم بياني مشابه لعرض فترات الثقة، الشكل (16.7) يتبني نفس الطريقة، ويستخدم الأمر (**twoway lfitci, stdf range(1979 2012**) ثم نقوم بوضع خط متصل (**connect**) فوق رسم بياني لدرجات الحرارة التي تم رصدها خلال الفترة 1979 – 2011 حيث تم تمثيلها بمثلثات مجوفة (**msymbol(Th)**) وشكل انتشار لدرجات الحرارة المتوقعة لسنة 2012 فقط مع علامة مربع (**msymbol(S)**) وتم إضافة نص يوضح القيم المتوقعة الرقمية، وحدود الثقة (*tempNhat, tempNlo, tempNhi*) لسنة 2012 تم نسخها من الجدول أعلاه) ويجب ملاحظة أن حدود فترة الثقة **lfitci, stdf** واسعة بما فيه الكفاية لتكون بمستوى 95% للمشاهدات، وليست مثل فترات الثقة التي كانت في الشكل (15.7) **lfit,stdp**.

```
.graph twoway lfitci tempN year, stdf
 lwidth(medthick)
 range(1979 2012)
 || connect tempN year, msymbol(Th)
 || scatter tempNhat year if year==2012,
 msymbol(S)
 || , ytitle("Annual temperature
 anomaly,`=char(186)'C")
 legend(off) xlabel(1980(5)2010) yline(0)
```

```
text(2.8 2007 "2012 predicted temp"
"1.79 (.84-2.73) `=char(186)'C")
title("Arctic temperature trend with 95% c.i.
for predictions"
, size(medlarge))
```



الشكل (16.7)

درجات الحرارة والمناطق المتجمدة والمتغيرات الأخرى من السلسلة الزمنية بملف البيانات *Arctic9.dta* هي نوع من البيانات التي تُنتج في العادة ارتباطاً ذاتياً autocorrelation أو ارتباطاً متسلسلاً بين قيم البيانات المتوالية. إذا كانت أخطاء الانحدار في الواقع مرتبطة ذاتياً، فإن هذا يعني أن المعادلات العادية للأخطاء المعيارية وفترات الثقة واختبارات الفرضيات – والتي تم استخدامها في هذا الجزء من الكتاب – يمكن أن تكون مضللة. وبالتالي فإن الباحثين في مجال بيانات السلاسل الزمنية ونماذجها يقومون بشكل دوري بفحص الارتباط الذاتي الباقي، ويقومون بتطبيق طرق انحدار خاصة للسلاسل الزمنية عند الحاجة.

طرق انحدار السلاسل الزمنية (الفصل 12) تتطلب بيانات يجب اعتبارها بيانات خاصة بسلاسل زمنية باستخدام الأمر **tsset**، هذا الأمر يحدد متغيرًا يستخدم كمؤشر للزمن.

```
.tsset year
 time variable: year, 1979 to 2012
 delta:1 year
```

بالنسبة لبيانات الأمر **tsset** هناك عدة طرق لفحص الارتباط الذاتي، إحدى هذه الطرق معروفة ولكنها أقل تفصيلاً، وهي اختبار دوربن واتسون Durbin-Watson.

```
.estat dwatson
 Durbin-Watson d-statistic( 2, 33) =  2.091689
```

العديد من أدلة الاستخدام تحتوي على جداول للبحث عن اختبار دوربن واتسون، مع تقدير 33 مشاهدة ومعلمتين 2. فإن القيمة المحسوبة 2.09 تقع أعلى من α=0.05 وهي الحد الأعلى للجدول. لذلك فإننا لا نرفض فرضية العدم القائلة بأنه ليس هناك ارتباط ذاتي موجب من الدرجة الأولى، وهذا يعتبر أمرًا جيدًا لصلاحية الشكل (15.7) والشكل (16.7)، ويعني أنه لايوجد تأكيد حول وجود الارتباط الذاتي عند استخدام فترات تباطؤ مثل سنتين أو ثلاث أو أربع سنوات ماضية.

هناك طريقة أكثر تفصيلاً لحساب الارتباط الذاتي لحساب مُعَامِلات الارتباط الذاتي للبواقي عند استخدام فترات تباطؤ مع اختبار التجميع التراكمي أو إحصائية ليجنج بوكس كيو Ljung-Box Q. هذا الاختبار يمكن القيام به من خلال تطبيق الأمر **corrgram** على بواقي النموذج (والتي تم تسميتها هنا *tempNres*).

```
.predict tempNres, resid
.corrgram tempNres
```

```
                                             -1       0       1 -1       0       1
 LAG      AC        PAC        Q      Prob>Q  [Autocorrelation] [Partial Autocor]

 1     -0.0803   -0.0826    .23248   0.6297
 2     -0.0920   -0.1081    .54749   0.7605
 3     -0.0494   -0.0746    .64146   0.8869
 4     -0.0249   -0.0461    .66619   0.9554
 5      0.2623    0.2818    3.5048   0.6227
 6     -0.1982   -0.2320    5.1858   0.5202
 7      0.1972    0.2678    6.9135   0.4379
 8     -0.0025    0.0217    6.9138   0.5460
 9     -0.1696   -0.2945    8.2974   0.5045
 10     0.1652    0.3323    9.6677   0.4701
 11    -0.2572   -0.5436     13.14   0.2843
 12     0.0647   -0.0919     13.37   0.3428
 13    -0.2205   -0.4464    16.177   0.2397
 14     0.0219   -0.0385    16.206   0.3009
```

اختبارات Q في جدول المخرجات أعلاه، ليست ذات معنوية عند فترات تباطؤ من 1 إلى 14 سنة ماضية، لذا فإن الأمر **corrgram** يتوافق بدرجة كبيرة مع نتائج الأمر **estat dwatson** التي توضح بأنه ليس هناك ارتباط ذاتي ذو معنوية إحصائية في بواقي نموذج درجات الحرارة. في هذا الجزء، فإن الاختبارات، وفترات الثقة لم يتم إخضاعها إلى أي اختبارات إضافية أخرى.

الرسومات البيانية التشخيصية : Diagnostic Graphs

برنامج ستاتا يوفر العديد من الرسومات البيانية المفيدة بغرض تشخيص نتائج نماذج الانحدار. القليل من هذه الرسومات تم شرحها في هذا الجزء. للحصول على قائمة بهذه الرسومات قم بطباعة الأمر **help regress postestimation**. الأمثلة في هذا الجزء سوف تكون بمثابة تجربة على نموذج الجليد في القطب الشمالي الذي يُظهِر بأن منطقة الجليد في شهر سبتمبر تم توقعها للفترة من *year* إلى *year*^{2} (بعد سنة *year* تم إجراء تمركز) مع درجات الحرارة غير العادية السنوية للهواء (*tempN*). بالنسبة لمتغير *year* تم إجراء تمركز له (*year0*) وتم تربيعه (*year02*) وتم حساب هذه المتغيرات سابقاً في هذا الفصل، ولكن تم إنشاء هذه المتغيرات من جديد بافتراض أنها غير موجودة، والمتغيرات التنبؤية الثلاثة معاً تشرح نحو 82% من التباين في المنطقة الجليدية.

```
.use C:\data\Arctic9.dta, clear
.gen year0 = year - 1995
.gen year02 = year0 ^2
.regress area year0 year02 tempN
```

```
      Source |       SS       df       MS              Number of obs =      33
-------------+------------------------------           F(  3,    29) =   49.72
       Model |  19.2134599     3  6.40448663           Prob > F      =  0.0000
    Residual |    3.735237    29  .128801276           R-squared     =  0.8372
-------------+------------------------------           Adj R-squared =  0.8204
       Total |  22.9486969    32  .717146777           Root MSE      =  .35889

------------------------------------------------------------------------------
        area |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       year0 |  -.0601115   .0111399    -5.40   0.000    -.0828951   -.0373279
      year02 |  -.0019336   .0008202    -2.36   0.025    -.0036111   -.0002562
       tempN |  -.2796799   .1538498    -1.82   0.079     -.594338    .0349783
       _cons |    5.24665   .1344514    39.02   0.000     4.971666    5.521634
------------------------------------------------------------------------------
```

كما شاهدنا في الجزء السابق، فإن اختبار بواقي الارتباط الذاتي معقول، حيث إن اختبار Q وجد أنه لايوجد ارتباط ذاتي ذو معنوية إحصائية عند استخدام فترة تباطؤ واحدة وحتى عشر فترات تباطؤ عند مقارنة بواقي كل سنة، ولكن الارتباط الذاتي يظهر عند استخدام فترات تباطؤ أكبر من 10، ولكن هذا الارتباط الذاتي من غير المحتمل أن يؤثر على النتائج.

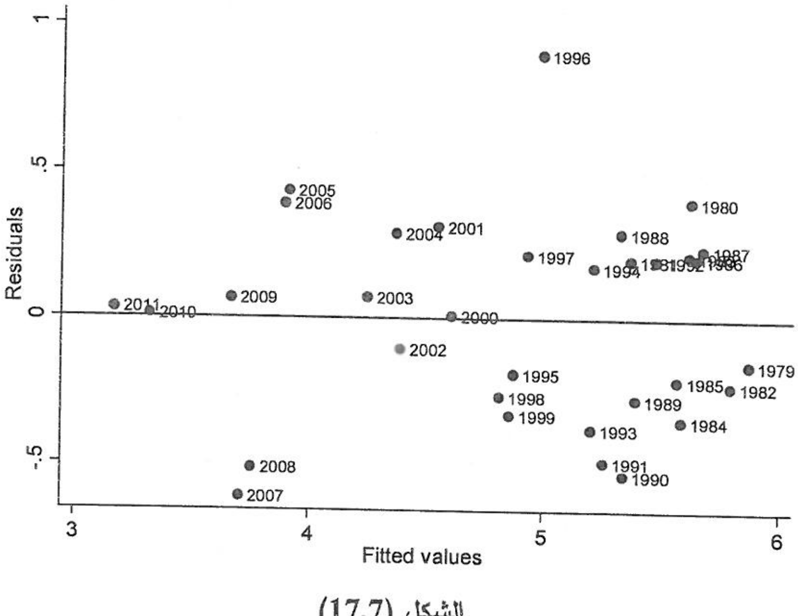
```
.predict areares2, resid
.corrgram areares2, lag(10)
```

```
                                          -1       0       1 -1       0       1
 LAG       AC       PAC      Q     Prob>Q  [Autocorrelation]  [Partial Autocor]
-------------------------------------------------------------------------------
 1        0.1140   0.1141   .46917  0.4934          |                 |
 2       -0.1826  -0.2003   1.7112  0.4250         -|                -|
 3       -0.3273  -0.2968   5.8358  0.1199        --|               --|
 4       -0.0554  -0.0157   5.9581  0.2023          |                 |
 5        0.0238  -0.1040   5.9816  0.3080          |                 |
 6       -0.1620  -0.4049   7.1046  0.3113         -|              ---|
 7       -0.1077  -0.1646     7.62  0.3673          |                -|
 8        0.2332   0.3384   10.132  0.2559          |-                |--
 9        0.3583   0.2410   16.309  0.0607          |--               |-
10       -0.0160  -0.2435   16.322  0.0908          |                -|
```

رسم بياني للبواقي مع القيم المناسبة يمكن إنشاؤه بواسطة حساب القيم المتوقعة، وإنشاء رسم بياني للمتغير *areares2* مع بقية المتغيرات الأخرى. الطريقة الأسرع للقيام بذلك تتم باستخدام الأمر **rvfplot**، المثال في الشكل

(17.7) يُضيف خطاً أفقياً لمستوى الصفر والمتوسط المتبقي، كما أنه يعطي وصفاً لنقاط البيانات لكل سنة *year* ، كما أن الشكل يوضح بأن قيمة متطرفة واحدة مع قيمة متبقية موجبة مرتفعة (1996)، ولكن لا توجد إشارة واضحة عن وجود مشاكل.

```
.rvfplot, yline(0) mlabel(year)
```



الشكل (17.7)

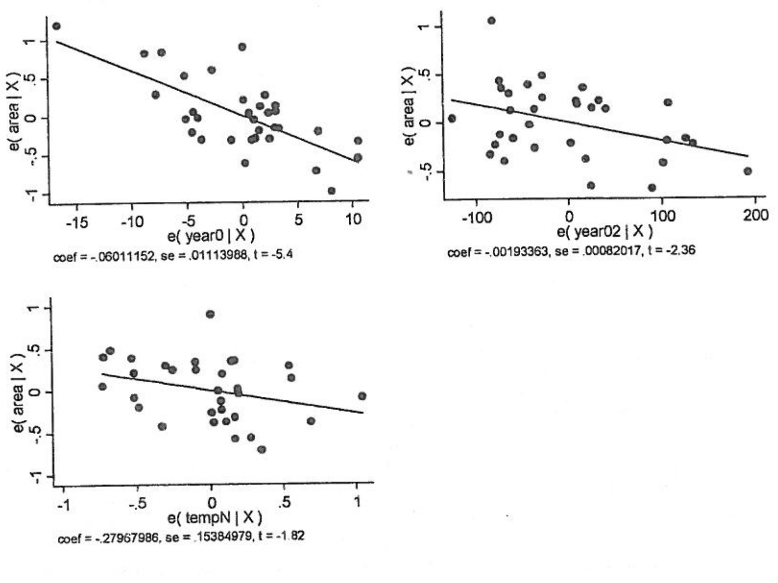
الرسم البياني للقيم المضافة تعتبر أدوات قيّمة، وتُعرف بأسماء عديدة منها: رسم بياني لتأثير الانحدار الجزئي، أو رسم بياني للبواقي الجزئية المعدلة، أو رسم بياني للمتغيرات المعدلة. وهذه الرسومات البيانية تصف العلاقة بين المتغير y ومتغير x واحد، وهي تعديل لتأثيرات متغيرات x الأخرى. وإذا كنا قد قمنا بحساب انحدار المتغير y على المتغيرين $x2, x3$ أو بالمثل حساب انحدار المتغير $x1$ على المتغيرين $x2, x3$ ثم نأخذ البواقي من كل انحدار ونقوم بإنشاء رسم بياني لهذه البواقي. وسوف نقوم بإنشاء رسم بياني لمتغير إضافي يوضح العلاقة بين المتغير y والمتغير $x1$ والمتغيرين المعدلين $x2, x3$، الأمر **avplot** يقوم بالحسابات الضرورية بشكل تلقائي. فمثلاً

إنشاء رسم بياني لمتغير إضافي للمتغير التنبؤي *tempN* تتم من خلال طباعة الأمر:

```
.avplot tempN
```

ولتسريع الحسابات أكثر، يمكننا القيام باستخدام الأمر **avplots** للحصول على مجموعة من الرسومات البيانية الصغيرة لكل متغير تنبؤي في تحليل الانحدار، الشكل (18.7) يعرض نتائج الانحدار للمتغير *area* على المتغيرات *year0, year02, tempN* الخطوط المرسومة في أشكال المتغير الإضافي لها ميل يساوي مُعَامِلات الانحدار الجزئية المقابلة. فعلى سبيل المثال، ميل الخط عند أسفل اليسار في الشكل (18.7) يساوي 0.2797- وهو بالضبط معامل المتغير *tempN* للثلاثة متغيرات التنبؤية في تحليل الانحدار.

```
.avplots
```



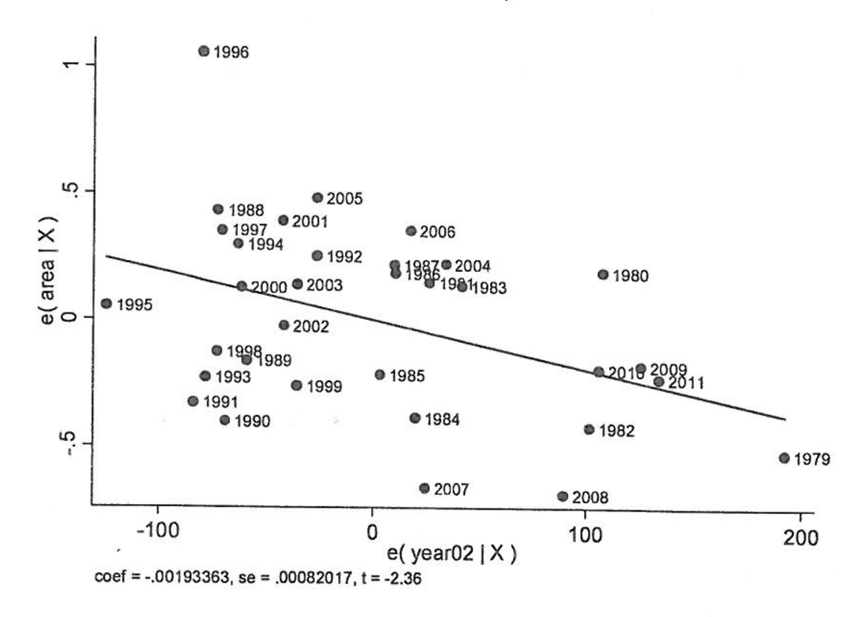
الشكل (18.7)

الرسومات البيانية للمتغير الإضافي تساعد في تحديد المشاهدات التي لها تأثير غير متكافئ في نموذج الانحدار. أما في الانحدار المتعدد، فإن إشارات التأثير تصبح أقل حدة، والمشاهدة التي بها مجموعة من القيم غير العادية في عدد من متغيرات x قد يكون تأثيرها مرتفعاً أو من المحتمل أن تؤثر على

الانحدار حتى ولو كانت إحدى قيم المتغير x هـي نفـسها غيـر عاديـة، المشاهدات ذات التأثير المرتفع تظهر في الرسم البياني للمتغير المضاف على شكل نقاط يمكن تمييزها عن باقي البيانات. وأغلب النقاط المتطرفـة التـي تظهر في الشكل (18.7) تظهر في مواقع ثابتة مع باقي البيانات الأخرى.

إحدى القيم المتطرفة تظهر في الرسم البياني في أعلى اليمين في الشكل (18.7) وتشير إلى تأثير محتمل أشد انحداراً (جعله أكثر سـلبية) لمعامـل المتغير *year02*، عندما نقوم بإنشاء رسم بياني لأحد المتغيـرات المـضافة باستخدام الأمر **avplot** وتوصيف نقاط بيانات الرسم، فإن سنة 1996 تظهر أنها هي تلك القيمة المتطرفة.

```
. avplot year02, mlabel(year)
```



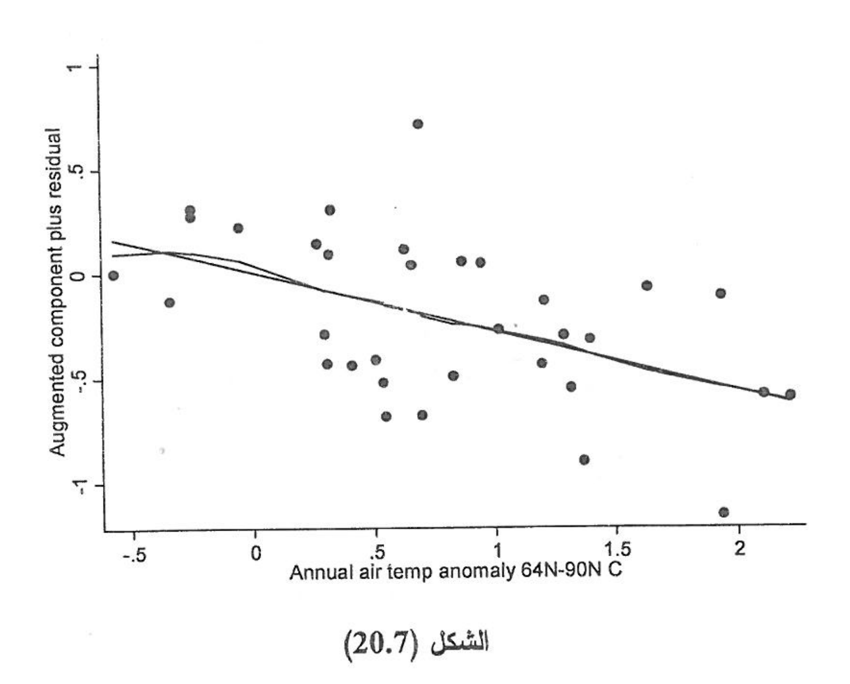
الشكل (19.7)

الرسومات البيانية للبواقي مضافاً إليها مكوناتها (والتـي تـم إنـشاؤها بواسطة الأمر **cprplot**) فإن المنحنى يأخذ شكلاً مختلفـاً. فالرسـم البيـاني للبواقي مضافاً إليها مكوناته للمتغير $x1$ يمثل بيانياً كل بـاقي مـضافاً إليـه مكوّنه، فإن نموذج التنبؤ للمتغير $x1$ هو:

$$e_i+b_1\,x1_i$$

مع قيم المتغير *x1*، مثل هذه الأشكال البيانية قد تساعد في تشخيص عدم الخطيّة، وتشير إلى نماذج عملية بديلة. الرسم البياني الفعّال للبواقي مضافاً إليها مكوناتها تعمل بطريقة ما أفضل، بالرغم من أن كلا النوعين في العادة يعطيان نتائج غير حاسمة (Mallows 1986)، الشكل (20.7) يعرض شكلاً بيانياً أكبر للبواقي مضافاً إليها مكوناتها الناتجة من انحدار المتغير *area* على المتغيرات *year0, year02, tempN*.

```
.acprplot tempN, lowess
```



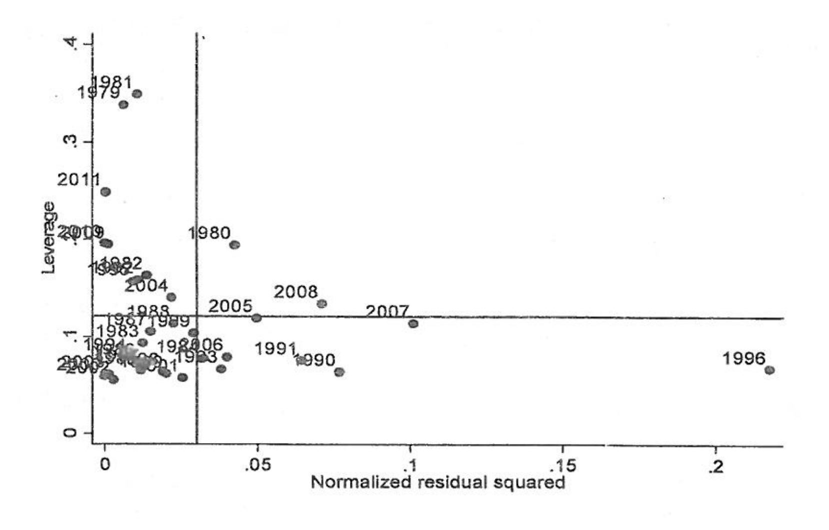
الشكل (20.7)

الخط المستقيم بالشكل (20.7) يتطابق مع نموذج الانحدار، بينما الخط المنحني يعكس تجانس المربعات الصغرى الموزونة المحلية، والتي يمكن أن تعرض لنا عدم الخطيّة أكثر. انخفاض المنحنى عند نهايته في الجانب الأيسر يمكن إهماله واعتباره مربعات صغرى موزونة اصطناعية، حيث هناك حالات قليلة تحدد اتجاهات هذا المنحنى (انظر الفصل 8). إذا كانت هناك أجزاء مركزية أكثر في منحنى المربعات الصغرى الموزونة تعرض نمطاً

منهجياً لانتقال المنحنى من نموذج الانحدار الخطي، فسوف يكون لدينا سبب للشك في كفاءة النموذج. في الشكل (20.7) قيم الوسيط للبواقي مضافاً إليها مكوناتها تتبع بشكل كبير نموذج الانحدار. الشكل يدعم النتائج التي تقول إن نموذج الانحدار الحالي يأخذ في الاعتبار عدم الخطية التي توجد في البيانات الخام ولا يترك أي بواقي.

كما يبدو أيضاً أن الرسم البياني يعرض تربيع البواقي مقابل التأثير (الخط المائل لمصفوفة التقدير) مع تربيع البواقي، الشكل (21.7) يعرض مثل هذا النوع من الرسم لانحدار المتغير *area*؛ ولتحديد القيم المتطرفة الفردية سوف نقوم بتوصيف العلامات بالرسم البياني بالسنوات *year*، الخيار **mlabsize(medsmall)** يحدد بأن أسماء العلامات سوف تكون متوسطة الحجم، وهي بطريقة ما أكبر من الحجم الافتراضي وهو الحجم الصغير (للحصول على قائمة بالخيارات الأخرى لحجم النص، قم بطباعة الأمر **help testsizestyle**)، الخيار **mlabpos(11)** يضع هذه الأسماء عند موقع الساعة 11 بالنسبة لرموز العلامات. أغلب السنوات في الرسم البياني تظهر متشابكة في أسفل اليسار في الشكل (21.7) ولكن سنة 1996 تظهر متطرفة في الخارج مرة أخرى.

```
.lvr2plot, mlabel(year) mlabsize(medsmall)
 mlabpos(11)
```



الشكل (21.7)

خطوط التأثير مع تربيع البواقي بالرسم البياني توضح متوسطات التأثير (الخط الأفقي)، وتربيع البواقي (الخط العمودي). التأثيريوضح احتمالية تأثير مشاهدة على الانحدار بناءً على مجموعة معينة من قيم المتغير x. أما قيم المتغير x المتطرفة أو المجموعات غير العادية فلها قيم تأثير مرتفعة، كما أن الارتفاع في تربيع البواقي يشير إلى أن مشاهدة ما مع قيمة للمتغير y تختلف كثيراً عن القيمة المتوقعة بواسطة نموذج الانحدار، سنة 1996 لها أكبر تربيع للبواقي، وهذا يشير إلى أن النموذج أقل تناسباً مع تلك السنة، ولكن القيم المكونة للنموذج وهي المتغيرات *tempN* و *year* تقع في المنتصف، وبالتالي فإن تأثير سنة 1996 أقل من المتوسط.

الرسومات البيانية التشخيصية والإحصائيات الأخرى تُشير إلى المشاهدات المؤثرة أو المشاهدات التي يُحتمل أن يكون لها تأثير ولكن هذه الرسومات لا توضح ما إذا كان يجب علينا استبعاد هذه المشاهدات، هذا يتطلب قرارًا موضوعياً بناءً على تقييم دقيق للبيانات، وتقييم للبحث بصفة عامة، وليس هناك تبرير موضوعي لاستبعاد سنة 1996 في مثال الجليد بالقطب الشمالي، ولكن يُفترض – للتوضيح فقط – أن نحاول إجراء ذلك على أي حال. وكما يُتوقع من التأثير المنخفض لسنة 1996 فإن إهمال هذه السنة يؤدي إلى اختلاف بسيط لنتائج تحليل انحدار *area*، مُعَامِلات المتغيران *year0, tempN* تبقى تقريباً هي نفسها مع تأثير يصبح ذا معنوية إحصائية للمتغير *tempN*. أما معامل المتغير *year02* يصبح أقرب للصفر، ولكنه مازال سالباً وذا معنوية إحصائية، R^2_a تزداد زيادة طفيفة من 0.82 إلى 0.85 عند استبعاد سنة 1996. في الحقيقة فإن هذه الاختلافات تعتبر طفيفة وليس لدينا سبب موضوعي لاستبعاد سنة 1996. النتائج توضح أنه من الأفضل الإبقاء على سنة 1996 في التحليل.

```
.regress area year0 year02 tempN if year!=1996
```

Source	SS	df	MS
Model	18.9699037	3	6.32330125
Residual	2.8640433	28	.102287261
Total	21.8339471	31	.704320873

Number of obs	=	32
F(3, 28)	=	61.82
Prob > F	=	0.0000
R-squared	=	0.8688
Adj R-squared	=	0.8548
Root MSE	=	.31982

area	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
year0	-.0602946	.0099275	-6.07	0.000	-.0806302	-.0399591
year02	-.0015288	.0007439	-2.05	0.049	-.0030527	-4.87e-06
tempN	-.2820721	.1371057	-2.06	0.049	-.5629203	-.0012239
_cons	5.182538	.1218136	42.54	0.000	4.933014	5.432062

دراسات كل مــن Cahmpers et al. (1983) وCook and Weisberg (1994) تعطي أمثلة أكثر تفصيلاً وشرحاً للرسومات البيانية التشخيصية، والطــرق البيانية الأخرى لتحليل البيانات.



الفصل الثامن

طرق الانحدار المتقدمة

Advanced Regression Methods

الفصل السابق ركّز على تحليل الانحدار بطريقة المربعات الصغرى (OLS). ولأسباب عدة فإن OLS إلى حد كبير تعتبر أكثر طرق تحليل الانحدار استخداماً. هذا الفصل يركز على مجموعة مختارة من طرق تحليل الانحدار الأخرى، والتي لها تطبيقات عدة، مع محاولة التركيز على التعقيدات التي لا توجد في طريقة OLS، وبالرغم من أن حساب الانحدار بالطرق الأخرى أكثر تركيزاً من OLS بسبب كثرة التحديات الرياضية، فإن هذه الطرق ليست صعبة الاستخدام ببرنامج ستاتا.

ليس هناك تسلسل معين لشرح مثل هذا الموضوع المتنوع، ولكن كل جزء في هذا الفصل تمت كتابته ليكون مستقلاً بذاته، لذا فإن القارئ يمكنه الانتقال بين هذه الأجزاء دون الحاجة إلى قراءة الجزء السابق، وتم العمل على جعل الأمثلة في هذا الفصل بسيطة، مع الإشارة إلى المصادر الأخرى التي يمكن أن يجد فيها القارئ تفاصيل أكثر.

مجموعة القوائم أدناه تغطي أغلب العمليات التي تمت مناقشتها في هذا الفصل، أحد هذه الموضوعات هو الانحدار غير الخطي، والذي يتطلب استخدام أسلوب الأوامر بدلاً من القوائم.

Graphics > Twoway graph (scatter, line etc.)
Statistics > Nonparametric analysis > Lowess smoothing
Statistics > Linear models and related > Other > Robust regression
Statistics > Linear models and related > Quantile regression
Statistics > Linear models and related > Box-Cox regression
Statistics > SEM (Structural Equation Modeling)

أمثلة عن الأوامر : Example Commands

```
.boxcox y x1 x2 x3, model(lhs)
```

يقوم بإيجاد أكبر التقديرات المتوقعة للمعلمة λ (لمدا) لتحويل كوكس – بوكس للمتغير $y^{(\lambda)}$ وهي دالة خطية للمتغيرات *x3, x2, x1* زائداً أخطاء التباين الثابت لجاوس. الخيار **model(lhs)** يحدد التحويل للطرف الأيسر للمتغير *y*، وهناك خيارات أخرى يمكن تحويلها للطرف الأيمن وهي متغيرات *x*، وذلك للتحكم بشكل أكثر في تفاصيل نموذج الانحدار؛ للحصول على تفاصيل عن كيفية بناء الأمر أعلاه وقائمة بخياراته قم بطباعة الأمر **help boxcox**، كما أن دليل المستخدم *Base Reference Manual* يشرح تفاصيل تقنية أكثر.

```
.graph twoway mband y x, bands(10) || scatter y x
```

يقوم بإنشاء رسم بياني لشكل انتشار المتغير *y* على المتغير *x* مع خط يصل نقاط الوسيط ببعضها (نقاط وسيط المتغير *x*، وسيط المتغير *y*) مع نطاقات أفقية بسُمك 10 نقاط لكل نطاق، وهذا أحد أنواع نطاقات الانحدار، وعند طباعة الخيار **mspline** بدلاً من الخيار **mband** في هذا الأمر، فإن ذلك سوف يؤدي إلى توصيل نقاط الوسيط بواسطة منحنى مائل يمر بكل النقاط بدلاً من خط متصل بين كل نقطة وأخرى.

```
.graph twoway lowess y x,bwidth(.4)|| scatter y x
```

يقوم بإنشاء رسم بياني لنقاط انتشار للمتغير *y* مع المتغير *x* ويقوم الخيار **bwidth(.4)** بحساب المربعات الصغرى الموزونة المحلية باستخدام عرض قدره 0.4 (40% من البيانات) وحتى يمكننا إنشاء قيم متجانسة كمتغير جديد، فإنه يجب استخدام الأوامر المتعلقة بالمربعات الصغرى الموزونة (يتناولها المثال التالي).

```
.lowess y x, bwidth(.3) gen(newvar)
```

يقوم بإنشاء رسم بياني للمربعات الصغرى الموزونة المتجانسة مع شكل انتشار المتغير *y* على المتغير *x* باستخدام نطاق سُمكه 0.3 (30% من البيانات)، القيم المتوقعة لهذا المنحنى سوف يتم تخزينها في متغير جديد باسم *newvar* يوجد بالأمر **lowess** خيار حفظ القيم المتوقعة، وهذا الخيار لا يمكن

القيام به باستخدام الأمر **graph twoway lowess**، وللحصول على مزيد من التفاصيل قم بطباعة الأمر **help lowess**.

```
.nl (y1={b1=1}*{b2=1}^x)
```

يقوم هذا الأمر باستخدام المربعات الصغرى غير الخطية المتعاقبة لتتناسب مع معلمتين في نموذج النمو الأُسّي $y=b_1b_2^x$. والمعلمتان اللتان يتم تقديرهما هما b1 وb2 محاطتان بالأقواس في الأمر أعلاه { } مع بداية مقترحة تبدأ بقيمة (1)، وبدلاً من كتابة النموذج في سطر الأمر، يمكننا توفير الوقت من خلال استخدام أحد أوامر النماذج المتوافرة في برنامج ستاتا أو كتابة برنامج جديد لتعريف نموذجنا الجديد. والمعلمتان الأسيّتان تحدثان لتكون واحدة من النماذج الأكثر شيوعاً ويتم تعريفهما بواسطة برنامج ستاتا باسم **exp2**، وبالتالي فإنه بإمكاننا القيام بنفس المهمة التي قام بها الأمر أعلاه وذلك من خلال طباعة الأمر:

```
.nl exp2: y x, init(b1 1 b2 1)
```

بعد طباعة الأمر **nl** قم باستخدام الأمر **predict** لاستخراج القيم المتوقعة أو البواقي.

```
.nl log4: y x, init(b0 5 b1 25 b2 .1 b3 50)
```

الأمر أعلاه يتناسب مع 4 معالم لنموذج النمو اللوغاريتمي **(log4)** ليكون على الشكل التالي:

$$y = b_0 + b_1/(1 + \exp(-b_2\,(x - b_3)))$$

قم بتحديد القيم المعلمية الأولية لعملية التقدير المتعاقبة عند b0=5، b1=25، b2 = 0.1، b3 =50؛ الخيار **log4** يشبه الخيار **exp2**، حيث إنه أحد النماذج غير الخطيّة الموجودة ببرنامج ستاتا.

```
.sem (y<- x1 x2 x3 x4) (x1<- x3 x4) (x2<- x3 x4)
```

هي نموذج لمعادلة هيكلية والتي يكون فيها *x1, x2, x3* يؤثر على *y, x1, x2* وهي متغيرات متداخلة كل منها يتم التأثير عليها بواسطة *x3, x4*.

```
.rreg y x1 x2 x3
```

يقوم بحساب الانحدار الموثوق للمتغير *y* مع ثلاثة متغيرات تنبؤية باستخدام المربعات الصغرى الموزونة التكرارية مع المعادلات الثنائية ومعادلات هوبر Huber عند مستوى 95% لكفاءة جاوس، بافتراض أن البيانات تم تجهيزها بطريقة مناسبة، فإن الأمر **rreg** يمكنه أيضاً حساب المتوسطات الموثوقة وفترات الثقة والاختلافات في اختبارات المتوسطات والتباين أو التغاير.

```
.rreg y x1 x2 x3, nolog tune(6) genwt(rweight)
 iterate(10)
```

يقوم بإجراء تحليل الانحدار الموثوق للمتغير *y* مع ثلاثة متغيرات تنبؤية، الخيارات المعروضة أعلاه تطلب من برنامج ستاتا عدم طباعة سجل المحاولات. قم باستخدام ثابت ضبط مقداره 6 (والذي يقلل من القيم المتطرفة بقوة أكثر من الوضع الافتراضي 7)، ولإنشاء متغير (عشوائياً يسمى *rweight*) يحتوي على الأوزان الموثوقة للتكرار النهائي لكل مشاهدة ويحدد الحد الأعلى للتكرار ليكون 10.

```
.qreg y x1 x2 x3
```

يقوم بحساب الانحدار الربيعي، والذي يُعرف كذلك باسم أقل قيمة مطلقة (LAV) أو باسم الانحدار المعدّل L1 للمتغير *y* على ثلاثة متغيرات تنبؤية، والوضع الافتراضي أن نماذج الأمر **qreg** تقوم بحساب الانحدار عند الربيع 0.5 (الوسيط تقريباً) للمتغير *y* كدالة خطيّة للمتغيرات التنبؤية، وبالتالي يعطي الانحدار الوسيط.

```
.qreg y x1 x2 x3, quantile(.25)
```

يقوم بحساب الانحدار الربيعي عند الربيع 0.25 (الربيع الأول) للمتغير *y* كدالة خطيّة للمتغيرات *x1, x2, x3*.

تجانس المربعات الصغرى المرجحة المحلية : Lowess Smoothing

تم الإشارة إلى تجانس المربعات الصغرى المرجحة (lowess) سابقاً في أكثر من موضوع من هذا الكتاب بدون شرح العديد من النقاط. تجانس المربعات الصغرى المرجحة هو أداة مفيدة للانحدار اللامعلمي، وعموماً فإن

طرق الانحدار اللامعلمية لا تحدد معادلة انحدار واضحة ولا تتطلب من التحليل مقدماً أن يحدد شكلاً عملياً للعلاقة. وبدلاً من ذلك، فإن هذه الطرق تساعد على اكتشاف البيانات بطريقة أكثر وضوحاً. هذا الإجراء يمكنه الكشف عن نتائج مثيرة أو غير متوقعة.

الأمر **lowess** والأمر **graph twoway lowess** يمكنهما إجراء التجانس (تجانس شكل الانتشار المرجح المحلي)، الأمر **lowess** مع الخيار **generate** يمكنه حفظ القيم المتوقعة. الأمر **graph twoway lowess** يتميز بالبساطة ويتبع نفس التركيب المعتاد الذي سبق استخدامه في هذا الكتاب، وله إمكانية دمج أكثر من شكل بياني في شكل واحد كما حدث سابقاً في مجموعة أوامر **graph twoway**. وكمثال بسيط سوف نقوم بإجراء رسم بياني لدرجات الحرارة العالمية غير العادية باستخدام مجموعة البيانات بالملف *global3.dta* والتي سبق استخدامها في الفصل (2).

```
.use C:\data\global3.dta, clear
.describe
```

```
Contains data from C:\data\global3.dta
  obs:         1,584                          Global climate
 vars:             5                          4 Jul 2012 11:21
 size:        20,592
```

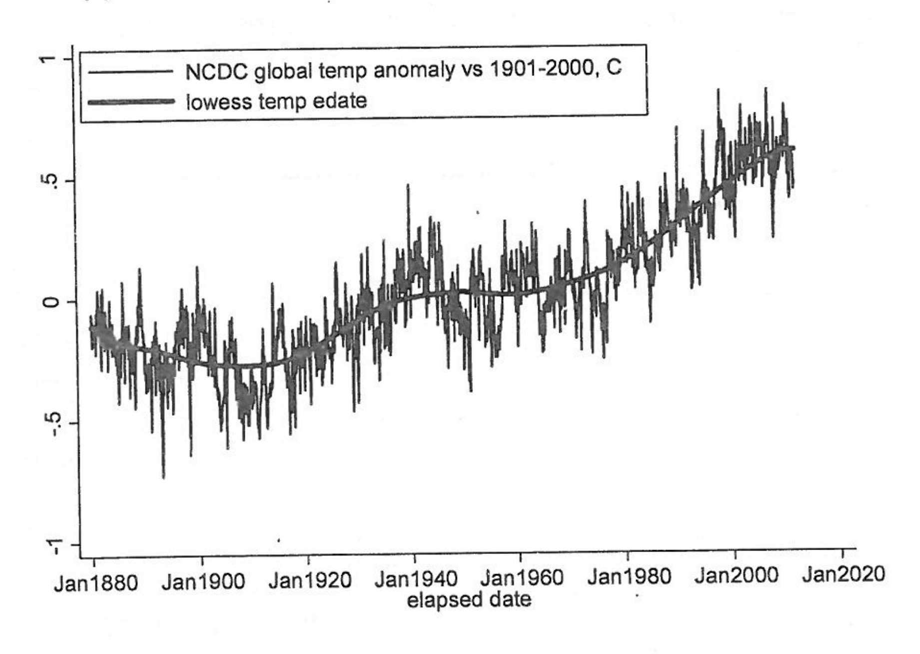
variable name	storage type	display format	value label	variable label
year	int	%8.0g		Year
month	byte	%8.0g		Month
edate	int	%tdmCY		elapsed date
temp	float	%9.0g		NCDC global temp anomaly vs 1901-2000, C
mei	float	%9.0g		Multivariate ENSO Index

```
Sorted by:  year  month
```

درجات الحرارة العالمية غير العادية في الفترة من 1880 إلى 2011 توضح تبايناً ملحوظاً، من شهر لآخر ومن سنة لأخرى، ومن المستحيل ملاحظة ما إذا كان التغير المناخي على المدى الطويل يتجه نحو ازدياد الحرارة أو البرودة أو سيبقى كما هو، تجانس المربعات الصغرى الموزونة

يساعد في ملاحظة التغير في المدى الطويل الذي يكمن في التقلبات الشهرية، الشكل (1.8) يعرض خطاً بيانياً لدرجات الحرارة والتواريخ الماضية (**twoway line *temp edate***) ثم يتم تركيب هذا الرسم على منحنى تجانس المربعات الصغرى المرجحة مع نطاق بسُمك 0.3 وخط أكثر سُمكاً لتوضيحه (**lowess *temp edate,* bw(.3) lwidth(thick)**)، وللحصول على معلومات عن الخيارات الأخرى للخطوط البيانية قم بطباعة الأمر **help linewidthstyle**.

```
.graph twoway line temp edate
    || lowess temp edate, bw(.3) lwidth(thick)
    || , legend(position(11) ring(0) rows(2))
```



الشكل (1.8)

منحنى lowess في الشكل (1.8) يعرض بوضوح مراحل ازدياد درجات الحرارة في فترة مبكرة من القرن العشرين (خصوصاً خلال الفترة 1920–1940) والانخفاض في درجات الحرارة في منتصف القرن، والازدياد السريع في درجات الحرارة (ما بعد سنة 1970) تتضمن نمطاً مستمرًا

للتغيرات في المناخ العالمي، نفس المراحل تظهر كذلك في البيانات المحلية في تواريخ ذوبان الجليد لبحيرة وينيبيسودكي Lake Winnipesaukee والذي يظهر في الشكل (26.3) في الفصل (3).

Lowess يتوقع تجانس قيم المتغير y لعدد n من المشاهدات الناتج من عدد من تحليلات الانحدار الموزونة n، بافتراض أن k يمثل نصف النطاق مع التقريب إلى أقرب عدد صحيح. لكل متغير y_i قيمة متجانسة y_i^s يمكن الحصول عليها بواسطة الانحدار الموزون الذي يتضمن هذه المشاهدات الموجودة في المدى i=max(1,i-k) وحتى i = min (i+k,n) حيث إن المشاهدة التي ترتيبها jth في المدى المحدد لها وزن w_j وذلك حسب دالة ثلاثية التكعيب:

$$w_i = (1-|u_j|^3)^3$$

حيث إن:

$$u_j = (x_i - x_j)/\Delta$$

Δ عبارة عن اختصار للمسافة بين x_i وأبعد قيمة في الفترة، الأوزان تساوي 1 لكل x_i=x_j ولكن هذه الأوزان قد تنخفض إلى الصفر عند حدود الفترة، انظر دراسة .Chambers et al (1983) أو دراسة Cleveland (1993) للحصول على تفاصيل وأمثلة عن طرق تجانس المربعات الصغرى المرجحة.

خيارات الأمر **lowess** تتضمن التالي:

mean وذلك لحساب المتوسط المتجانس، الوضع الافتراضي هو وضع خط تجانس المربعات الصغرى.

noweight التجانس غير الموزون، والوضع الافتراضي هو حساب دالة كليفلاند Cleveland الموزونة ثلاثية التكعيب.

bwidth() تحديد المدى للمجموعات الفرعية المتجانسة لمدى من المشاهدات n × التي تم استخدامها للتجانس – باستثناء نقاط النهاية التي تكون أصغر – وفترات عدم التأكد المستخدمة، الوضع الافتراضي هو **bwidth(.8)**

logit تحويل القيم المتجانسة إلى قيم لوغاريتمات.

adjust تعديلات المتوسط للقيم المتجانسة لتساري متوسط المتغير *y* الأصلي مثل **logit, adjust** وهي مفيدة مع المتغير الثنائي *y*.

gen(*newvar*) إنشاء متغير جديد اسمه *newvar* يحتوي على القيم المتجانسة للمتغير *y*.

nograph يمنع هذا الخيار عرض الرسم البياني ضمن النتائج.

addplot() تتم إضافة رسومات بيانية أخرى للرسم البياني الموجود، ولمزيد من التفاصيل قم بطباعة **help addplot option**.

lineopts() يؤثر على عرض خط التجانس، ولمزيد من التفاصيل قم بطباعة **help cline options**، وحيث إن هذا يتطلب *n* من الانحدارات الموزونة لذلك فإن تجانس المربعات الصغرى المرجحة قد يستغرق وقتاً عند حسابه للعينات الكبيرة.

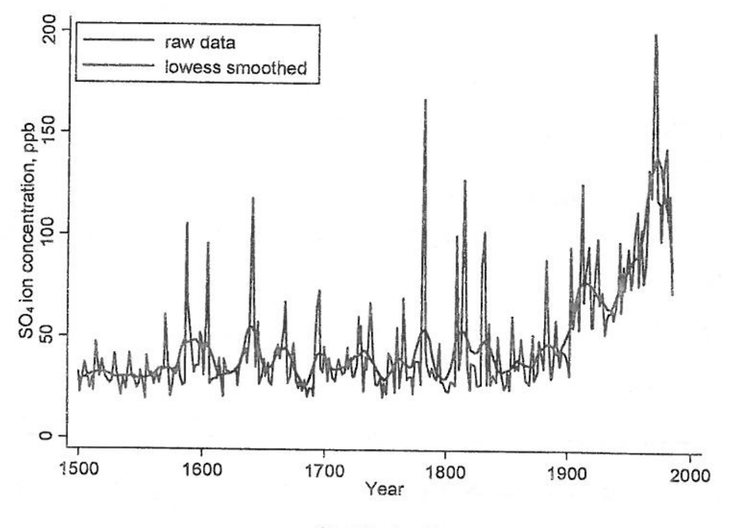
مثل طرق التجانس الأخرى (أو أي نموذج)، فإن تجانس المربعات الصغرى المرجحة يقوم بتقسيم البيانات إلى أجزاء: جزء متجانس مثل المنحنى السميك في الشكل (1.8) وجزء تقريبي وهو في اليسار بعد طرح التجانس من البيانات. وعادة فإن الجزء التقريبي يحتوي على معلومات مفيدة أيضاً، ولشرح ذلك سوف نقوم بالانتقال إلى مجموعات بيانات عن طبقات الجو العليا خلال عدة قرون زمنية وهي تتضمن قياسات تم الحصول عليها من الطبقات الجليدية لجليد جرين لاند (GISP2) وهذه البيانات تم شرحها في دراسة Mayewski, Holdsworth and colleagues (1993) ودراسة Mayewski, Meeker and colleagues (1993)، وقد قام الباحثان بالحصول على عينات من هذا الجليد وتحليلها كيميائياً، وهذه البيانات تمثل أكثر من 100,000 سنة من التغير المناخي، هذه البيانات موجودة بالملف *Greenland_sulfate.dta* وتتضمن جزءًا بسيطاً من المعلومات عن تركيز الأملاح الكبريتية غير البحرية، ومؤشر عن كثافة التوزيع القطبية.

```
.use C:\data\Greenland_sulfate.dta, clear
.describe
```

```
Contains data from C:\data\Greenland_sulfate.dta
  obs:          271                          Greenland ice core sulfate & PCI, 1500-1985
                                               (Mayewski 1993)
 vars:            3                          2 Jul 2012 06:11
 size:        4,878
-----------------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-----------------------------------------------------------------------------------------
year            int    %ty                    Year
sulfate         double %10.0g                 SO4 ion concentration, ppb
PCI             double %6.0g                  Polar Circulation Intensity
-----------------------------------------------------------------------------------------
Sorted by:  year
```

وللحصول على تفاصيل أكثر عن السلسلة الزمنية التي تحتـوي علـى 271 نقطة، يجب القيام بتجانس مع نطاق ضيق يُعادل 5% من العينة. الشكل (2.8) يعرض رسماً بيانياً لنتائج المتغير *sulfate* وهو يمثل تركيز الأمـلاح الكبريتية غير البحرية.

```
.graph twoway line sulfate year
    || lowess sulfate year, bwidth(.05)
    lwidth(medthick)
    || , ytitle("SO{subscript:4} ion
    concentration, ppb")
    legend(label(1 "raw data") label(2 "lowess
    smoothed")
    position(11) ring(0) rows(2))
```



الشكل (2.8)

الأملاح الكبريتية غير البحرية (SO_4) وصلت للمنطقة الجليدية بجرينلاند بعد أن تم إضافتها في طبقات الجو عن طريق البراكين أو الوقود المستخرج من الفحم والنفط. كلا المنحنيين (المتجانس والخام) بالشكل (2.8) يعطيان هذه المعلومات. المنحنى المتجانس يعرض تذبذباً بأعلى من المتوسط بقليل من 1500 وحتى بدايات سنوات 1800، أما بعد 1900 فإن الوقود المستخرج قد دفع منحنى التجانس إلى الارتفاع مع انخفاض مؤقت بعد سنة 1929 (فترة الكساد العظيم) وبدايات السبعينيات (التأثير الناتج من القانون الأمريكي للهواء النظيف في سنة 1970، وحصار النفط العربي في سنة 1973، والارتفاع الكبير في أسعار النفط الذي صاحب حصار النفط)، أغلب الرؤوس المدببة في منحنى البيانات الخام تم تحديدها مع ثورات البراكين المعروفة مثل بركان هيكلا Hekla في أيسلاند (1970) أو بركان كاتماي Katmai في آلاسكا (1912).

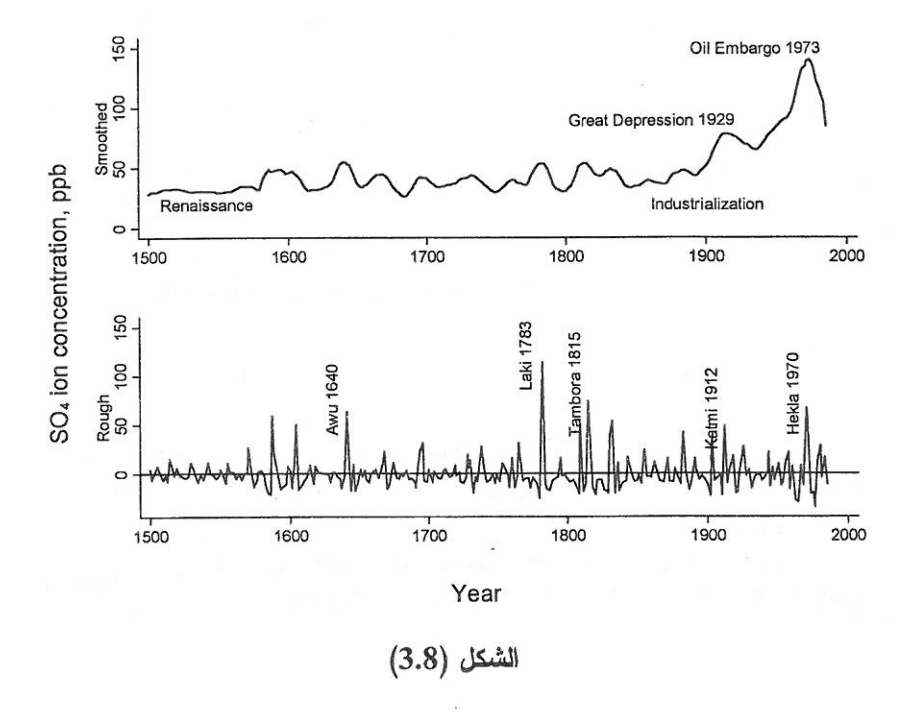
بعد تجانس بيانات السلاسل الزمنية، فإنه من المفيد دراسة التجانس والتقريب (البواقي) للسلاسل بشكل منفصل، أدناه قمنا باستخدام الأمر **lowess** لتحديد متغيرين جديدين: الأول للقيم المتجانسة للأملاح الكبريتية (*smooth*)، والثاني للبواقي أو القيم التقريبية (*rough*) ويتم حسابها بطرح القيم المتجانسة من البيانات الخام.

```
.lowess sulfate year, bwidth(.05) gen(smooth)
.label variable smooth "SO4 ion concentration
 (smoothed)"
.gen rough = sulfate - smooth
.label variable rough "SO4 ion concentration
 (rough)"
```

الشكل (3.8) يُقارن بين السلاسل الزمنية للمتغير *smooth* والمتغير *rough* في زوج من الرسومات البيانية التي تم وضع شروح لها باستخدام الخيارات **text()**، لاحظ استخدام الخيارات **saving()** في بداية أول أمرين لإنشاء الرسم البياني، هذان الشكلان تم وضعهما في شكل واحد باستخدام الأمر **graph combine**، في الشكل الموحّد تم استخدام محور أفقي واحد وتم وضع عنوان لهذا

المحور باستخدام ("b1 b1title("Year، في بداية الأمريــشير أول عنــوان فــي الأسفل، الشكل الموحّد لا يتعرّف على المحور x والمحور y ولكن يتعرّف علــى العنوان أسفل الشكل (b1 وb2)، وبالنسبة للعناوين فــي يــسار الــشكل (l1 وl2) والعناوين في أعلى الشكل (t1 وt2) وفي يمين الــشكل (r1 وr2)، فــي الــشكل (3.8) عنوان المحور العمودي y تم إنشاؤه في بداية اليــسار وذلــك باســتخدام الخيار ("l1title("SO{subscript:4} ion concentration, ppb.

```
.graph twoway line smooth year,ylabel(0(50)150)
 xtitle("")
 lwidth(medthick)lcolor(maroon)ytitle("Smoothed")
 text(20 1540 "Renaissance") text(20 1900
 "Industrialization")
 text(90 1860 "Great Depression 1929")
 text(150 1935 "Oil Embargo 1973")
 saving(fig08_03a.gph, replace)
.graph twoway line rough year, ylabel(0(50)150)
 xtitle("")
 ytitle("Rough") text(75 1630 "Awu 1640",
 orientation(vertical))
 text(120 1770 "Laki 1783",
 orientation(vertical))
 text(90 1805 "Tambora 1815",
 orientation(vertical))
 text(65 1902 "Katmai 1912",
 orientation(vertical))
 text(80 1960 "Hekla 1970",
 orientation(vertical))
 yline(0) saving(fig08_03b.gph, replace)
.graph combine fig08_03a.gph fig08_03b.gph,
 rows(2)b1title("Year")
 l1title("SO{subscript:4} ion concentration,
 ppb")
```



الشكل (3.8)

الانحدار الموثوق : Robust Regression

يقوم الأمر **regress** والأمر **anova** بحساب انحدار المربعات الصغرى العادي (OLS)، القبول الكبير لـ (OLS) كان يرجع جزئياً إلى مميزاته النظرية هذا في حالة الحصول على البيانات الصحيحة، وإذا كانت الأخطاء موزعة توزيعاً طبيعياً مستقلاً ومحدداً فإن OLS يعتبر أكثر كفاءة من أي مُقدّر آخر غير متحيّز. الجانب الآخر من هذه العبارة هو إذا كانت الأخطاء غير موزعة توزيعاً طبيعياً مستقلاً ومحدداً، فإن المقدّرات الأخرى غير المتحيّزة قد تكون أفضل من OLS، وفي الحقيقة فإن كفاءة OLS تنخفض بسرعة في مواجهة توزيعات خطأ ذات منحنى توزيع طبيعي ذو ذيل طويل (قيم متطرفة – ذات نزعة)، مثل هذه التوزيعات مازالت موجودة بشكل كبير في العديد من المجالات والحقول العلمية.

تميل OLS إلى تتبع القيم المتطرفة على حساب باقي القيم الموجودة بالعينة. وخلال المدى الطويل، فإن هذا يقود إلى تباين كبير من عينة إلى عينة أو عدم الكفاءة عند احتواء العينات على قيم متطرفة، لذا فإن الانحدار الموثوق يهدف إلى تعظيم كفاءة OLS في حالة وجود بيانات مثالية وزيادة هذه الكفاءة في الأوضاع المعقدة مثل وجود أخطاء غير طبيعية. الانحدار الموثوق يشمل تقنيات متنوعة، كل تقنية لها مميزاتها وعيوبها عند التعامل مع البيانات المعقدة. في هذا الجزء من الكتاب، سوف يتم شرح نوعين من الانحدار الموثوق هما **rreg** و**qreg** وسوف تتم مقارنتهما مع **regress**.

لاحظنا في الفصل (7) الانخفاض الواضح والحاد في المنطقة المنخفضة لحجم الجليد في المناطق القطبية خلال الفترة 1979-2011، ولكن ماذا حول جليد البحر في المنطقة القطبية الجنوبية؟ النمط الجغرافي والفصلي للمنطقة القطبية الجنوبية يختلف عما هو عليه في المنطقة القطبية الشمالية. ففي وسط القطب الشمالي المحيط مُحاط باليابسة والتي أصبحت أكبر بعد الزيادة الكبيرة في ذوبان الجليد في السنوات الأخيرة، حيث إن اليابسة زادت مساحتها بأكثر من 3 ملايين كيلومتر مربع أو 4 ملايين كيلومتر مربع في المنطقة التي بها 15% على الأقل من الجليد خلال فصل الصيف، ومن ناحية أخرى، فإن المنطقة القطبية الجنوبية عبارة عن قطعة من اليابسة محاطة بالمحيط، بينما جليد البحر في المنطقة القطبية الشمالية يمتد إلى القطب الشمالي إلا أن الجليد في القطب الجنوبي أقل امتداداً للقطب الجنوبي، ونسبة كبيرة من الجليد في المنطقة القطبية الجنوبية تذوب كل صيف، وعند بلوغ مستوى الحد الأدنى السنوي في شهر فبراير، فإن الجليد في المنطقة القطبية الجنوبية ينخفض إلى نحو 2 مليون كيلومتر مربع، ويمتد إلى أقل من 3 ملايين كيلومتر مربع؛ ملف البيانات *Antarctic2.dta* يحتوي على بيانات عن متوسط امتداد الجليد خلال شهر فبراير (البيانات من دراسة Milke and Heygster 2009) وتغطي الفترة من 1972 إلى 2011، كما تتضمن البيانات

درجات الهواء السنوية غير العادية للمنطقة القطبية الجنوبية، وهي تمتد من 64 إلى 90 درجة جنوباً، والتي تم تقديرها بواسطة وكالة ناسا.

```
.describe

Contains data from C:\data\Antarctic2.dta
  obs:            39                          Antarctic February mean sea ice 1973-2011 Milke
                                                & Heygster 2009
 vars:             4                          2 Jul 2012 06:11
 size:           429

              storage  display     value
variable name   type   format      label      variable label

year            int    %8.0g                  Year
yeargrp         byte   %9.0g       dec        1972-1999 v. 2000-2011
extentS         float  %9.0g                  SH sea ice extent, million km^2
tempS           float  %9.0g                  Annual air temp anomaly 64S-90S C

Sorted by:  year
```

هل الحد الأدنى لامتداد جليد البحر في المنطقة القطبية الجنوبية يميل نحو الازدياد أو النقص؟. انحدار OLS أوضح علاقة ضعيفة وليست ذات معنوية إحصائية لميله نحو التناقص ($p = 0.125$).

```
.regress extentS year

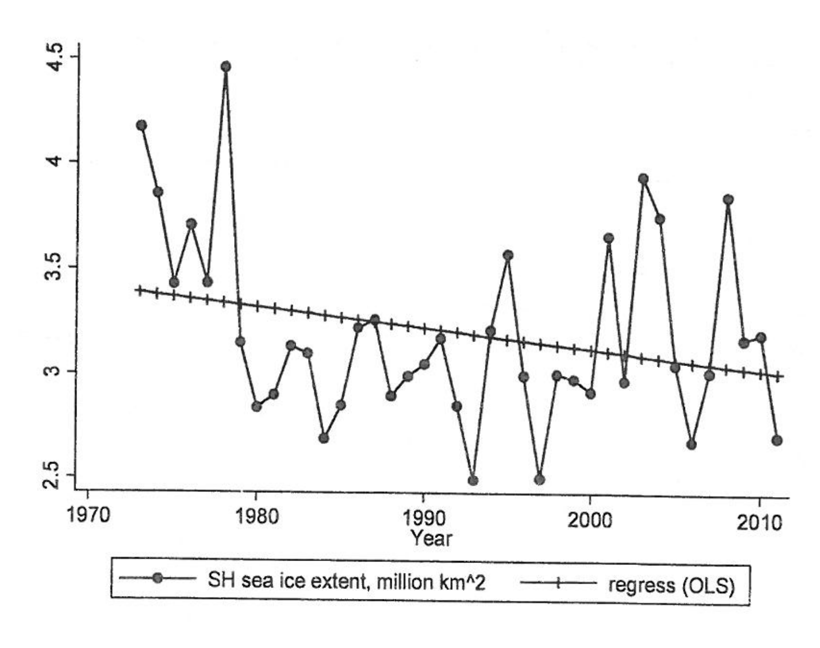
      Source |       SS       df       MS              Number of obs =      39
-------------+------------------------------           F(  1,    37) =    2.46
       Model |  .480675664     1  .480675664           Prob > F      =  0.1253
    Residual |  7.22814772    37  .195355344           R-squared     =  0.0624
-------------+------------------------------           Adj R-squared =  0.0370
       Total |  7.70882338    38  .202863773           Root MSE      =  .44199

------------------------------------------------------------------------------
     extentS |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |  -.0098642   .0062885    -1.57   0.125     -.022606    .0028776
       _cons |   22.84955   12.52695     1.82   0.076    -2.532471    48.23156
------------------------------------------------------------------------------
```

تناقص درجات الحرارة في المنطقة القطبية الشمالية كان واضحاً في الرسومات البيانية (الشكل 9.7 والشكل 12.7). الرسم البياني لتناقص درجات الحرارة في المنطقة القطبية الجنوبية في الشكل (4.8) لا يوضح اتجاهاً معيّناً، حيث إننا لا نرى أي مشاكل إحصائية محتملة، القيم المرتفعة

للمتغير *extentS* في سنتي 1972 و1977، هي فترة كانت فيهـا مـشاهدات الأقمار الصناعية أقل وضوحاً، وربما تؤثر هـذه المـشاهدات علـى خـط الانحدار، وتسبب في ميله السالب الضعيف.

```
.predict exthat1
.label variable exthat1 "regress (OLS)"
.graph twoway connectextentS exthat1 year,
 msymbol(O +)
```



الشكل (4.8)

الانحدار الموثوق يقاوم تأثير القيم المتطرفة، ممـا يجعـل الانحـدار الموثوق يتناسب مع الفحص السريع لمعرفة ما إذا كانت القيم المتطرفة لهـا تأثير غير مناسب على نتائج OLS، الأمر **rreg** يقوم بحساب تحليل الانحدار الموثوق، وعند تطبيقه على جليد البحر للمنطقة القطبية الجنوبية اتضح بـأن هناك انخفاضاً – ولكنه أقل حدة – وليس ذا معنوية.

```
.rreg extentS year
```

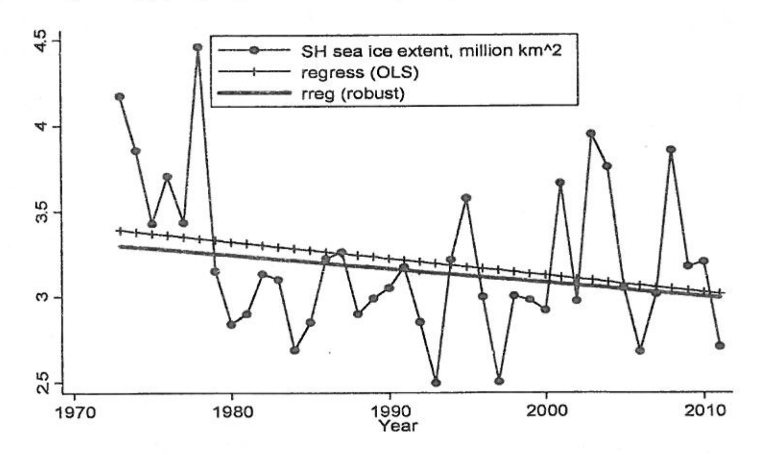
```
   Huber iteration 1:  maximum differente in weights = .61809736
   Huber iteration 2:  maximum difference in weights = .1095464
   Huber iteration 3:  maximum difference in weights = .0319613
Biweight iteration 4:  maximum difference in weights = .23592582
Biweight iteration 5:  maximum difference in weights = .08565176
Biweight iteration 6:  maximum difference in weights = .02170203
Biweight iteration 7:  maximum difference in weights = .00318406

Robust regression                                Number of obs =        39
                                                 F(  1,     37) =      1.62
                                                 Prob > F       =    0.2105
```

extentS	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
year	-.0080677	.0063304	-1.27	0.210	-.0208943	.0047588
_cons	19.21456	12.61029	1.52	0.136	-6.336321	44.76544

بعد تطبيق الأمر **rreg** يمكن للأمر **predict** – بشكل عادي – الحصول على القيم المتوقعة. وتمثل هذه القيم المتوقعة بيانياً (والتي تسمى هنا *exthat2*)، الشكل (5.8) يقارن بيانياً بين خط OLS وخط OLS الموثوق.

```
.predict exthat2
.label variable exthat2 "rreg (robust)"
.graph twoway connect extentS exthat1 exthat2
 year,
 msymbol(O + i) lwidth(medium medium thick)
 legend(ring(0) position(12) col(1))
```



الشكل (5.8)

الأمر **rreg** يعمل بواسطة المربعات الصغرى الموزونة التكرارية (IRLS)، التكرار الأول للأمر **rreg** يبدأ مع OLS، وأي مشاهدات لها تأثير

كبير مثل أن قيم مسافة كوك لها Cook's D أكبر من 1. فإن هذه المـــشاهدات سوف يتم استبعادها تلقائياً بعد الخطوة الأولى، ثم بعد ذلك يتم حساب الأوزان لكل مشاهدة باستخدام دالة هوبر Huber function (وهذه الدالة تقلل المشاهدات التي لها بواق كبيرة) ثم يتم حساب المربعات الصغرى الموزونة. بعد تكـــرار حساب المربعات الصغرى الموزونة عدة مرات، فإن دالة الترجيح أو الوزن سوف تنتقل إلى وزن توكي الثنائي Tukey (تم الإشارة إلى هذا الـــوزن فـــي دراسة Li 1985)، ويتحول إلى توزيع جـــاوس بكفـــاءة 95%. (لمزيـــد مـــن التفاصيل انظر دراسة Hamilton (1992a))، الأمر **rreg** يقوم بتقدير الأخطـــاء المعيارية ويختبر الفرضيات مستخدماً طريقة قيم وهمية لا تعتمد على فرضية التوزيع الطبيعي (انظر دراسة Street, Carroll and Ruppert 1988).

الأمر **rreg** والأمر **regress** كلاهمـــا ينتمـــي لمجموعـــة مُقـــدِّرات *M*- (احتمالية الحد الأقصى maximum). البديل الإحصائي الاستراتيجي لهـــذين الأمرين يسمى تقدير *L*-، ويتناسب بشكل كبير مع ربيعات y بدلاً من توقعات أو متوسطات التقدير نفسه. فمثلاً يمكننا إنشاء نمـــوذج كيـــف أن الوســـيط (الربيع 0.5) للمتغير y يتغير مع المتغير x، الأمر **qreg** (وهو نـــوع مـــن تقدير-*L1*) يقوم بحساب الانحدار الربيعي، الأمر **qreg** يشبه الأمر **rreg** حيث إنه يقاوم تأثير القيم المتطرفة. وعموماً فإن الأمر **qreg** يميل ليكون أقل كفاءة أو به أخطاء معيارية مرتفعة مقارنة مع الأمر **rreg**. وهذا هو الوضع هنـــا حيث إن الأمر **qreg** وجد ميلاً بسيطاً ولكن أخطاء معيارية أكبـــر، الـــشكل (6.8) يقارن بيانياً بين النماذج الخطيّة الثلاثة.

```
.qreg extentS year

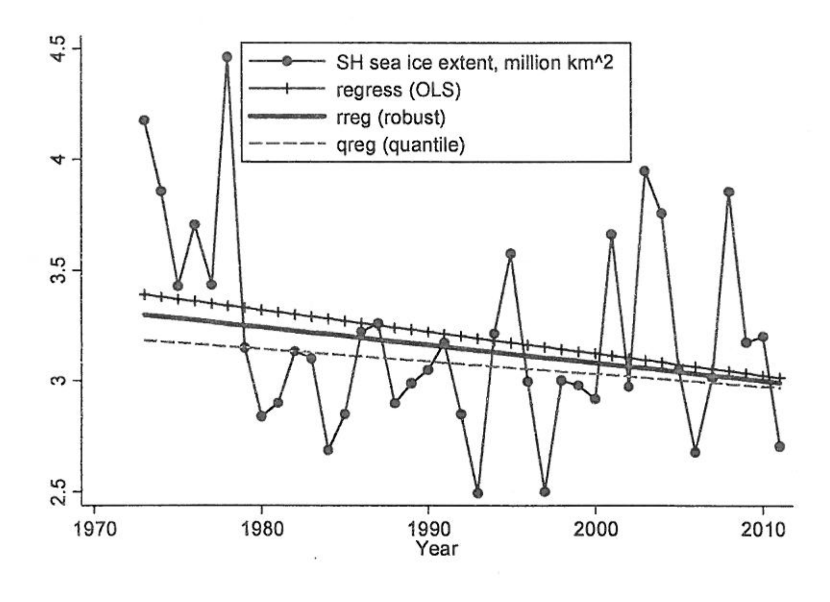
Iteration  1:  WLS sum of weighted deviations =  12.830409

Iteration  1: sum of abs. weighted deviations =  14.282429
Iteration  2: sum of abs. weighted deviations =  12.595795
Iteration  3: sum of abs. weighted deviations =   12.59334

Median regression                                Number of obs =        39
  Raw sum of deviations 12.99536 (about 3.0977242)
  Min sum of deviations 12.59334                 Pseudo R2     =    0.0309
```

extentS	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
year	-.0056349	.0054759	-1.03	0.310	-.0167301	.0054604
_cons	14.29918	10.9048	1.31	0.198	-7.796049	36.3944

```
.predict exthat3
.label variable exthat3 "qreg (quantile)"
.graph twoway connect extentS exthat1 exthat2 exthat3 year,
 msymbol(O + i i) lwidth(medium medium thick medthick)
 lpattern(solid solid solid dash) legend(ring(0) position(12) col(1))
.graph twoway connect extentS exthat1 exthat2 exthat3 year,
 msymbol(O + i i) lwidth(medium medium thick medthick)
 lpattern(solid solid solid dash) legend(ring(0) position(12) col(1))
.graph twoway connect extentS exthat1 exthat2 exthat3 year,
 msymbol(O + i i) lwidth(medium medium thick medthick)
 lpattern(solid solid solid dash) legend(ring(0) position(12) col(1))
.graph twoway connect extentS exthat1 exthat2
 exthat3 year,
msymbol(O + i i) lwidth(medium medium thick
 medthick)
lpattern(solid solid solid dash)
legend(ring(0) position(12) col(1))
```



الشكل (6.8)

الوضع الافتراضي هو أن يقوم الأمر qreg بحساب الانحدار الوسيط، ولكن هناك قدرات عامة أخرى لهذا الأمر، حيث إن هذا الأمر له القدرة على

حساب النماذج الخطيّة لأي ربيع من ربيعات المتغير *y* ، وليس فقط الوسيط (الربيع 0.5). فمثلاً الأمر أدناه وجد أن الربيع الثالث (الربيع 0.75) للمتغير *extentS* انخفض بطريقة ما بشكل أكبر من انخفاض الوسيط خـلال الفتـرة الزمنية. الميل للربيع 0.75 يساوي 0.0149− وهذا يعني انخفاضاً بمقـدار 14,900 كيلومتر مربع في السنة مقارنة مع 5,600 كم^2 في السنة للربيـع 0.5 أو الوسيط. وعموماً فإن هذه النتائج ليست ذات معنوية إحصائية.

.qreg *extentS year*, quant(.75)

```
Iteration  1:  WLS sum of weighted deviations =  12.774128

Iteration  1: sum of abs. weighted deviations =  12.877446
Iteration  2: sum of abs. weighted deviations =  12.702146
Iteration  3: sum of abs. weighted deviations =  12.323326
Iteration  4: sum of abs. weighted deviations =  12.284196

.75 Quantile regression                           Number of obs =        39
  Raw sum of deviations 12.70912 (about 3.4314537)
  Min sum of deviations  12.2842                  Pseudo R2     =    0.0334
```

extentS	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
year	-.0149048	.0179872	-0.83	0.413	-.0513504	.0215408
_cons	33.15648	35.83128	0.93	0.361	-39.4446	105.7576

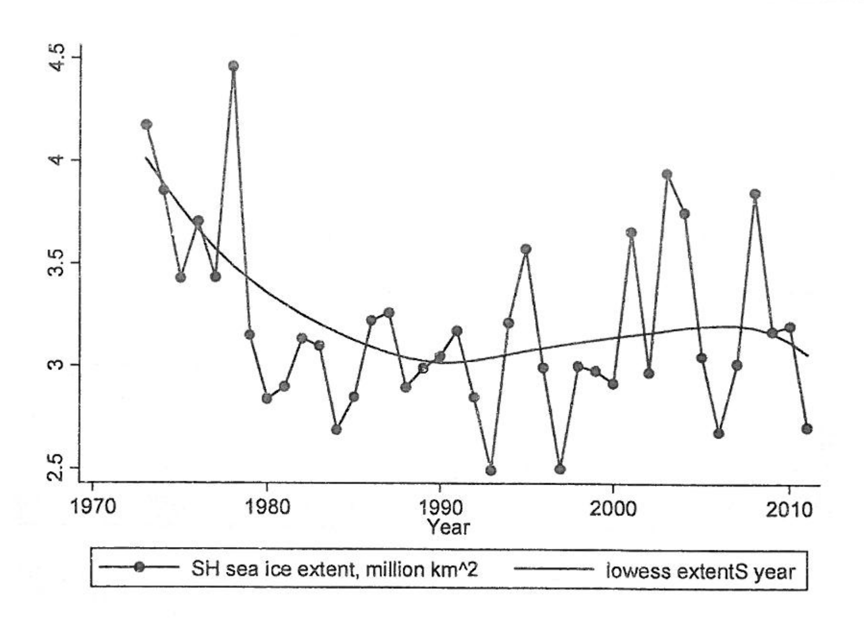
بافتراض ثبات تباين الخطأ، فإن ميول خطوط الربيعـات 0.25 و0.75 يُفترض أن تكون هي نفسها، وبالتالي فإن الأمر **qreg** يمكنه فحص اختلاف التباين heteroskedasticity أو أنواع بسيطة من عدم الاعتدال.

مشابهاً للأمر **regress**، فإن الأمر **rreg** والأمر **qreg** يمكنهما العمل مع المتغيرات المحوّلة (المحوّلة كلوغاريتمات أو ربيعات) والعمل مع أي عـدد من المتغيرات التنبؤية بما في ذلك المتغيرات الوهمية أو التفاعلات، الكفاءة وسهولة الاستخدام بشكل خاص تجعل الأمر **rreg** ذا قيمة للقيام بفحص عام وسريع لمعرفة ما إذا كانت نتائج الأمر **regress** تأثرت بالقيم المتطرفـة أو أخطاء التوزيعات غير الطبيعية، وإذا تم تطبيق الأمر **rreg** والأمر **regress**

على نفس النموذج، سوف نحصل على نفس النتائج تقريباً، ويمكننا تحديد استنتاجات بدرجة ثقة أكبر. إما إذا لم تتفق نتائج الأمر **rreg** مع نتائج الأمر **regress**، فإن ذلك يعتبر بمثابة علامة تنبيه بأن الاستنتاجات غير مستقرة، وتحتاج إلى مزيد من التحليل لمعرفة أسباب الاختلاف، وتحديد كيفية التعامل مع هذه المشاكل الإحصائية.

الاختلافات بين نتائج الأوامر **regress, rreg, qreg** ليست كبيرة في مثال جليد البحر بالمنطقة القطبية الجنوبية. نتائج الأوامر الثلاثة جميعها تتفق بأن هناك علاقة ضعيفة وليست ذات دلالة إحصائية للميل نحو الانخفاض. الأمر **regress** يعطي ميلاً أقل حدة في هذا الاتجاه، وذلك بسبب تأثره بالسنوات الأولى التي كانت قيمها مرتفعة. لهذا السبب فإن نماذج الأمر **rreg** أو الأمر **qreg** قد تكون مفضّلة عن غيرها، ولكن يمكننا السؤال عما إذا كان نموذج خطيّ معقول في بداية التحليل، انحدار تجانس المربعات الصغرى المرجحة والذي لا يفترض أي شكل عملي محدد يُعتبر أداة للإجابة عن الأسئلة من هذا النوع. عند تطبيق انحدار التجانس على بيانات جليد البحر بالمنطقة القطبية الشمالية (النتائج لن يتم عرضها هنا) فإن انحدار التجانس أظهر منحنى متشابهاً تماماً لذلك الذي أنتجه النموذج الربيعي في الفصل السابق بالشكل (12.7)، أما تطبيقه على جليد البحر بالمنطقة القطبية الجنوبية في الشكل (7.8) التالي فإن النتائج تُشير إلى أنه لاشيء يتشابه مع النموذج الخطي أو النموذج الربيعي وبدلاً من ذلك فإن الانحدار المتجانس يعطي شرحاً نوعياً للانخفاض الأوّلي والارتفاع الذي عقب ذلك ثم الانخفاض في السنوات الأخيرة. الانخفاض الأولي هو الوحيد الذي يظهر كبيراً، وتفسير ذلك هو وجود القدرة المحدودة للأقمار الصناعية في تسجيل التغيرات في تلك الفترة، وربما قد يكون نمط التغير في المدى الطويل سوف يكون أكثر وضوحاً في السنوات القادمة، ولكن هذا النمط ليس واضحاً من البيانات الموجودة لدينا.

```
.graph twoway connect extentS year || lowess
 extentS year
```



الشكل (7.8)

تطبيقات أخرى للأمر rreg والأمر qreg :

Further rreg and qreg Applications

الجزء السابق، عرض تطبيقات مبسّطة للأمر **rreg** والأمر **qreg**. هــذه الأوامر يمكن استخدامها أيضاً بعدة طرق من الطرق السهلة إلـــى الطـــرق الأكثر تعقيداً. فمثلاً للحصول على فترة ثقة 90% لمتوسط متغير واحد مثل درجة حرارة الهواء في المنطقة القطبية الجنوبية (*tempS*) يمكننا طباعة أمر فترة الثقة **ci**.

```
.ci tempS, level(90)
```

Variable	Obs	Mean	Std. Err.	[90% Conf.	Interval]
tempS	39	.3351282	.065531	.2246459	.4456105

أو بدلاً من ذلك، يمكننا الحصول على نفس المتوسط، وفترة الثقة مـــن خلال تحليل انحدار بدون إدخال متغيرات *x*، الخيار **nohead** يمنـــع ظهـــور جدول الانحدار حيث ليس له حاجة هنا.

```
.regress tempS, nohead level(90)
```

tempS	Coef.	Std. Err.	t	P>|t|	[90% Conf.	Interval]
_cons	.3351282	.065531	5.11	0.000	.2246459	.4456105

وبالمثل، يمكننا الحصول على متوسط موثوق مع فترة ثقة 90%، الخيار **nolog** يمنع ظهور سجل التكرار الموثوق وذلك توفيراً للمساحة.

```
.rreg tempS, nolog level(90)
```

```
Robust regression                               Number of obs =        39
                                                F(  0,     38) =      0.00
                                                Prob > F       =         .
```

tempS	Coef.	Std. Err.	t	P>|t|	[90% Conf.	Interval]
_cons	.318898	.0707103	4.51	0.000	.1996837	.4381124

الأمر qreg: يمكن استخدامه بنفس الطريقة للحصول على فترات ثقة تقريبية لوسيط أو أكثر، مع ملاحظة أن الربيع 0.5 الذي تم الحصول عليه عن طريق الأمر **qreg** قد لايكون مساوياً للوسيط. نظرياً الربيع 0.5 والوسيط متساويان. وعملياً فإن الربيعات هي قيم تقريبية من القيم الحقيقية للعينة، حيث إن الوسيط يتم حسابه عن طريق حساب متوسط القيمتين المركزيتين في البيانات في حالة أن مجموعة فرعية تحتوي على عدد زوجي من المشاهدات. لذا فإن وسيط العينة والربيع 0.5 يمكن أن يكونا مختلفين بطريقة لا تؤثر كثيراً على تفسير النموذج.

```
.qreg tempS, nolog level(90)
```

```
Median regression                               Number of obs =        39
  Raw sum of deviations     12.63 (about .28)
  Min sum of deviations     12.63               Pseudo R2     =    0.0000
```

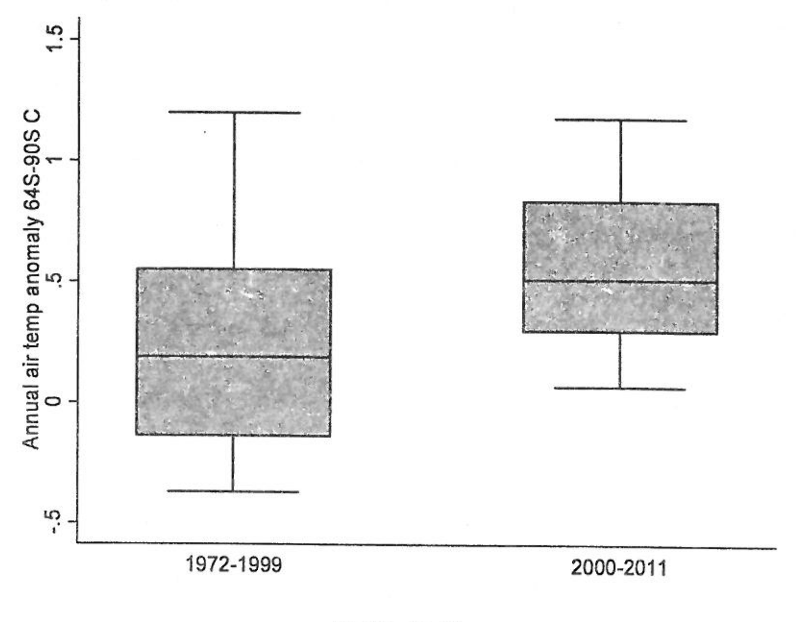
tempS	Coef.	Std. Err.	t	P>|t|	[90% Conf.	Interval]
_cons	.28	.075718	3.70	0.001	.1523428	.4076572

المتوسط الموثوق أقل بقليل من المتوسط العادي (0.319 مقابل 0.335) والربيع 0.5 يساوي (0.28) هو أيضاً أقل، مما يشير إلى وجود سنوات قليلة

شهدت ارتفاعاً في درجات الحرارة، مما أدى إلى زيادة قيمة المتوسط. في الأوامر أعلاه، الخيار level() يحدد درجة الثقة المرغوبة، وإذا قمنا بإهمال هذا الخيار، فإن ستاتا يقوم تلقائياً بحساب فترة ثقة 95%.

ولمقارنة المتوسطين يقوم المحللون باستخدام اختبار *t* للعينتين (**ttest**) أو تحليل التباين في اتجاه واحد (**oneway** أو **anova**). وكما شاهدنا سابقاً، يمكننا حساب اختبارات التعادل (وتؤدي إلى الحصول على إحصائية *F* وإحصائية *t*) وذلك من خلال تحليل انحدار المتغير الرقمي على متغير وهمي. المتغير الوهمي *yeargrp* والذي يساوي 0 للفترة 1972–1999 ويساوي 1 للفترة 2000–2011 يقدم تفسيراً لذلك. الشكل (8.8) يعرض بيانياً درجات الحرارة غير العادية لهاتين المجموعتين.

```
.graph box tempS, over(yeargrp)
```



الشكل (8.8)

تحليل الانحدار يؤكد الانطباع الذي حصلنا عليه من الشكل (8.8)، حيث إن السنوات الأخيرة شهدت ارتفاعاً ملحوظاً في درجات الحرارة، وهذا

الارتفاع ذو معنوية إحصائية ($p = 0.026$)، متوسط درجات الحرارة غير العادية في المنطقة القطبية الجنوبية يساوي 0.239 درجة مئوية للفترة 1992–1999، أما للفترة 2000–2011 فإن المتوسط أعلى بـ 0.312 درجة مئوية ($0.239 + 0.312 = 0.551$ درجة مئوية)، وتجدر الملاحظة أنه حتى مع ارتفاع درجات الحرارة نصف درجة مئوية، فإن المنطقة القطبية الجنوبية مازالت مكاناً بارداً جداً. ودرجات الحرارة سوف تستمر تحت الصفر لسنوات حول منطقة القطب الجنوبي.

```
.regress tempS yeargrp, nohead
```

tempS	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
yeargrp	.3115741	.1344639	2.32	0.026	.0391244	.5840237
_cons	.2392593	.0745871	3.21	0.003	.0881314	.3903871

ولكن، هل يمكننا الثقة في هذه النتيجة؟ الأمر **rreg** أوضح أن المتوسطات الموثوقة أكثر اختلافاً بقيمة قدرها 0.329 درجة مئوية، هذا الاختلاف ذو معنوية إحصائية ($p = 0.023$).

```
.rreg tempS yeargrp, nolog

Robust regression                               Number of obs =       39
                                                F(  1,     37) =     5.60
                                                Prob > F      =   0.0233
```

tempS	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
yeargrp	.3290205	.1390096	2.37	0.023	.0473602	.6106807
_cons	.2155881	.0771087	2.80	0.008	.0593511	.3718251

الأمر **qreg** وجد أن الربيع 0.5 يختلف أيضاً، حيث إنه يساوي 0.38 درجة مئوية. هذا الاختلاف ليس ذا معنوية إحصائية ($p = 0.082$)، وحيث إن هذه الاختلافات ليست ذات معنوية إحصائية بسبب زيادة الأخطاء المعيارية في تحليل الانحدار الذي تم بالأمر **qreg**، فإن هذا يؤدي إلى انخفاض إحصائية t، أن زيادة الأخطاء المعيارية تعكس انخفاض كفاءة الأمر **qreg**.

```
.qreg tempS yeargrp, nolog
```

```
Median regression                                     Number of obs =        39
  Raw sum of deviations     12.63 (about .28)
  Min sum of deviations     11.54                     Pseudo R2     =    0.0863

------------------------------------------------------------------------------
       tempS |     Coef.   Std. Err.       t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     yeargrp |       .38    .137707     2.76   0.009     .1009791    .6590209
       _cons |       .19    .079176     2.40   0.022     .0295742    .3504257
------------------------------------------------------------------------------
```

مع تأثير الترميز، وشروط التفاعل المناسبة، يمكن للأمر **regress** تكرار تحليل ANOVA بالضبط. حيث إن استخدام الأمر **regress** مع اتباعه بأوامر الاختبار المناسبة، سوف نحصل على نفس نتائج R^2 واختبار F والتي يمكن الحصول عليها باستخدام الأمر **anova**. القيم المتوقعة التي يتم الحصول عليها من مثل تحليلات الانحدار هذه تساوي متوسطات المجموعة. ويمكن للأمر **rreg** إجراء تحليلات متماثلة لاختبار الاختلافات في المتوسطات الموثوقة بدلاً من المتوسطات العادية. ويمكن استخدام الأمر **qreg** بنفس النمط للحصول على احتمال ثالث لاختبار الاختلافات في قيم الوسيط. كل هذا يتيح لنا صياغة نماذج متشابهة مع تحليل التباين المتعدد *N*-way ANOVA أو التغاير ANCOVA مع تضمن الربيع 0.5 أو تقدير قيم التباين بدلاً من المتوسطات المعتادة.

بغض النظر عن شكل توزيع الخطأ، فإن OLS تظل مُقدِّرًا غير متحيّز، وفي الأمد الطويل، فإن تقديرات OLS يُفترض أن تتمركز في قيم معلمية صحيحة. هذه الحالة لا تنطبق على أغلب المقدِّرات الموثوقة، وإذا لم تكن الأخطاء متماثلة، فإن خط الوسيط الذي تم إنشاؤه بالأمر **qreg** أو الخط ثنائي الوزن الذي تم إنشاؤه بواسطة الأمر **rreg** نظرياً تتزامن مع خط *y* المتوقع والذي تم تقديره بواسطة الأمر **regress**، وطالما أن التواء الأخطاء يعكس جزءًا صغيرًا من توزيعها، فإن الأمر **rreg** قد يكون أقل تحيّزاً. ولكن عندما يكون الالتواء في التوزيع بالكامل، فإن الأمر **rreg** سوف يقلل وزن جانب واحد من النموذج، مما يؤدي إلى تقديرات تقاطع-*y* متحيزة بشكل ملحوظ. تقديرات ميل الأمر **rreg** تظل غير متحيزة بالرغم من الالتواء في توزيعات

الأخطاء. لذا هناك مفاضلة بين استخدام الأمر **rreg** أو مُقدِّر مشابه مع الأخطاء الملتوية. نحن نخاطر بالوقوع في تقديرات متحيزة لتقاطع -y ولكن يمكن توقّع تقديرات دقيقة و غير متحيزة لمُعَامِلات الانحدار الأخرى. في العديد من التطبيقات مثل المُعَامِلات هي أكثر إثارة للاهتمام إلى حد كبير من تقاطع -y مما يجعل المفاضلة أمرًا مفيدًا جداً. وعموماً فإن اختبارات F و t ليست مثل OLS فهي لا تفترض وجود الأخطاء الطبيعية.

الانحدار غير الخطي – 1 : 1 – Nonlinear Regression

تحويل المتغيرات يسمح بإنشاء بعض العلاقات غير الخطية باستخدام تقنيات مألوفة للنماذج الخطية الحقيقية. ومن ناحية أخرى، فإن النماذج الخطية الحقيقية تتطلب مستوى آخر من التقنيات المتناسبة. الأمر **nl** يقوم بحساب الانحدار غير الخطي بواسطة المربعات الصغرى المتعاقبة. هذا الجزء يشرح مع الأمثلة التوضيحية للبيانات الموجودة بالملف *nonlin.dta*.

```
.use C:\data\nonlin.dta, clear
.describe
```

```
Contains data from C:\data\nonlin.dta
  obs:           100                          Nonlinear model examples (artificial data)
 vars:             5                          2 Jul 2012 06:11
 size:         1,700

              storage  display     value
variable name   type   format      label      variable label

x               byte   %9.0g                  Independent variable
y1              float  %9.0g                  y1 = 10 * 1.03^x + e
y2              float  %9.0g                  y2 = 10 * (1 - .95^x) + e
y3              float  %9.0g                  y3 = 5 + 25/(1+exp(-.1*(x-50))) + e
y4              float  %9.0g                  y4 = 5 + 25*exp(-exp(-.1*(x-50))) + e

Sorted by:  x
```

بيانات الملف *nonlin.dta* تم إنشاؤها مع متغيرات y والتي تُعرِّف دوال غير خطيّة متنوعة للمتغير x زائداً أخطاء جاوس العشوائية، فمثلاً المتغير $y1$ يمثل عملية النمو الأسّي

$$y1 = 10 \times 1.03^x$$

تقدير هذه المعلمات من البيانات يمكن الحصول عليها باستخدام الأمر **nl**

$$yl = 11.20 \times 1.03^x$$

وهو قريب بدرجة كبيرة من النموذج الصحيح.

```
.nl (y1 = {b1=1}*{b2=1}^x)

(obs = 100)

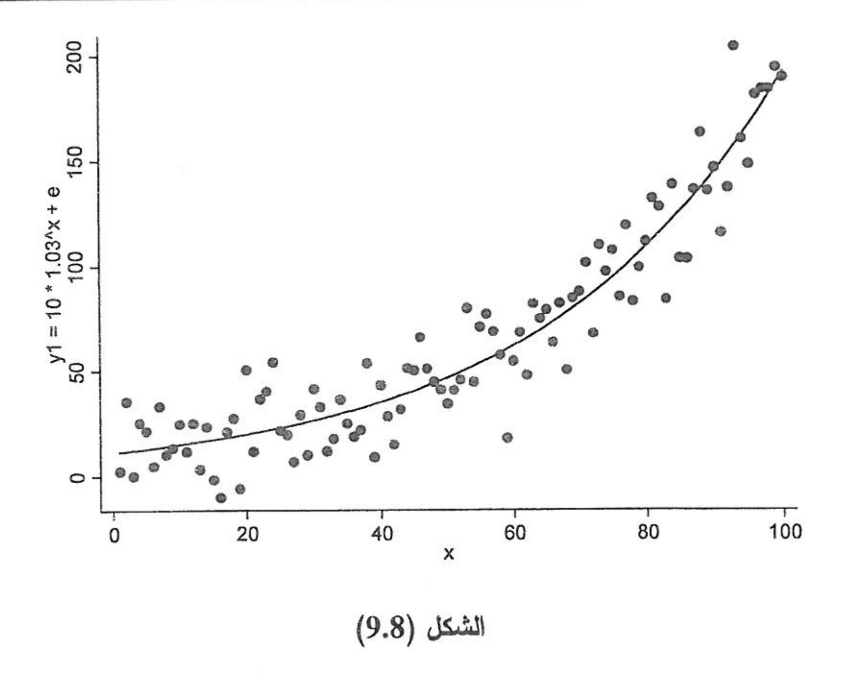
Iteration 0:  residual SS =  419135.4
Iteration 1:  residual SS =  416152.4
Iteration 2:  residual SS =  409107.7
Iteration 3:  residual SS =  348535.9
Iteration 4:  residual SS =  31488.48
Iteration 5:  residual SS =  27849.49
Iteration 6:  residual SS =  26139.18
Iteration 7:  residual SS =  26138.29
Iteration 8:  residual SS =  26138.29
Iteration 9:  residual SS =  26138.29

      Source |       SS       df       MS
-------------+------------------------------         Number of obs =        100
       Model |  667018.255     2  333509.128         R-squared     =     0.9623
    Residual |  26138.2933    98  266.717278         Adj R-squared =     0.9615
-------------+------------------------------         Root MSE      =   16.33148
       Total |  693156.549   100  6931.56549         Res. dev.     =   840.3864

------------------------------------------------------------------------------
          y1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         /b1 |   11.20416   1.146683     9.77   0.000     8.928602    13.47971
         /b2 |   1.028838   .0012404   829.41   0.000     1.026376    1.031299
------------------------------------------------------------------------------
```

الأمر **predict**: يقوم بحساب القيم المتوقعة والبواقي للنموذج غير الخطي الذي تم تقديره باستخدام الأمر **nl**. الشكل (9.8) يعرض رسماً بيانياً للقيم المتوقعة من المثال السابق، موضحاً التناسب الأقرب ما بين النموذج والبيانات (R^2=0.96).

```
.predict yhat1
.graph twoway scatter y1 x
      || line yhat1 x, sort
      || , legend(off) ytitle("y1 = 10 * 1.03^x
      + e")xtitle("x")
```



الشكل (9.8)

بدلاً من كتابة نموذج الأمر **nl**، يمكننا الحصول على نفس النتائج بطباعة الأمر التالي:

```
.nl exp2: y1 x
```

الخيار **exp2** في الأمر أعلاه، يقوم باستخدام برنامج اسمه *nlexp2.ado* والذي يُعرّف دالة النمو الأسّي ذات معلمتين. برنامج ستاتا يتضمن العديد من تلك البرامج لتطبيق الدوال التالية:

exp3 ثلاث معلمات أُسيّة: $y=b_0+b_1b_2^x$

exp2 معلمتان أسيّتان: $y= b_1b_2^x$

exp2a معلمتان أسيّتان سالبتان $y=b_1(1-b_2^x)$

log4 أربع معلمات منطقيّة، b_0 تمثل مستوى البداية و(b_0+b_1) مقارباً للحد الأعلى الذي يساوي $y=b_0+b_1/(1+\exp(-b_2(x-b_3)))$

log3 ثلاث معلمات منطقية تبدأ من الصفر وb_1 تقارب للحد الأعلى $y=b_1/(1+\exp(-b_2(x-b_3)))$

gom4 أربع معلمات لجومبرتز Gompertz، b_0 تمثـــل مـــستوى البدايـــة و(b_0+b_1) مقارب للحد الأعلى $y=b_0+b_1/\exp(-\exp(-b_2(x-b_3)))$

gom3 ثلاث معلمات لجومبرتز Gompertz، تبدأ من الصفر وb_1 تقـــارب للحد الأعلى $y=b_1\exp(-\exp(-b_2(x-b_3)))$

يمكن للمستخدمين كتابة برامج أخرى nl*function* خاصة بهم، كما يمكنك استخدام nlgom4.ado, nlexp3.ado أو الأمثلة الأخرى أعلاه، وللحصول علـــى تفاصيل وشروحات عن كيفية تحديد وتقدير النماذج قم بطباعة الأمر **help nl**.

الانحدار غير الخطي – 2 : 2 – Nonlinear Regression

بيانات الجليد لشهر سبتمبر في المنطقة القطبية الـــشمالية (*Arctic9.dta*) تعطي مثالاً حقيقياً. في الأمثلة السابقة، رأينا أن منطقة جليد البحر انخفضت في الفترة 1979–2011 وهي الفترة القريبة من مشاهدات الأقمار الصناعية. الرسومات البيانية التشخيصية توضح نموذجاً خطياً مناسباً ولكن بشكل سيء وذلك بسبب أن الانخفاض كان أسرع من الوضـــع الخطّـــي (الـــشكل 9.7 والشكل 11.7)، النموذج الآخر للمنطقة القطبية الجنوبية أفضل ويشرح اتجاه المشاهدات خلال سنة 2011 (الشكل 12.7)، إذا تم توقّع النموذج الربيعـــي لسنوات قليلة مقدماً، فإن اتجاهه يكون مستحيلاً فعلياً، حيث إنه سوف يصل للصفر بسرعة كبيرة، ويستمر ليكون سالباً. النماذج الفعلية تميـــل لإظهـــار انخفاض تدريجي حتى يصل هذا الانخفاض إلى الصفر (على سبيل المثـــال، انظر دراسة Wang and Overland 2009). النظير البسيط لمثل هذه النمـــاذج الفعلية قد تكون منحنى -S تماثلي، وهذا المنحنى مثل جـــومبرتز Gompertz بدلاً من الانخفاض السريع في النموذج الربيعي.

الأوامر أدناه تعمل على نموذج جومبرتز ثلاثي المعلمات لجليد البحـــر في المنطقة القطبية الشمالية. سوف نركز على المتغير *extent* (المناطق التي بها جليد نسبته 15% على الأقل) بدلاً من المتغير *area* (المناطق التي بهـــا

جليد 100%) والذي تم استخدامه سابقاً، وكما رأينا سابقاً في الشكل (14.3) فإن المتغيرين يسلكان نفس السلوك.

```
.use C:\data\Arctic9.dta, clear
.nl gom3: extent year, nolog
```

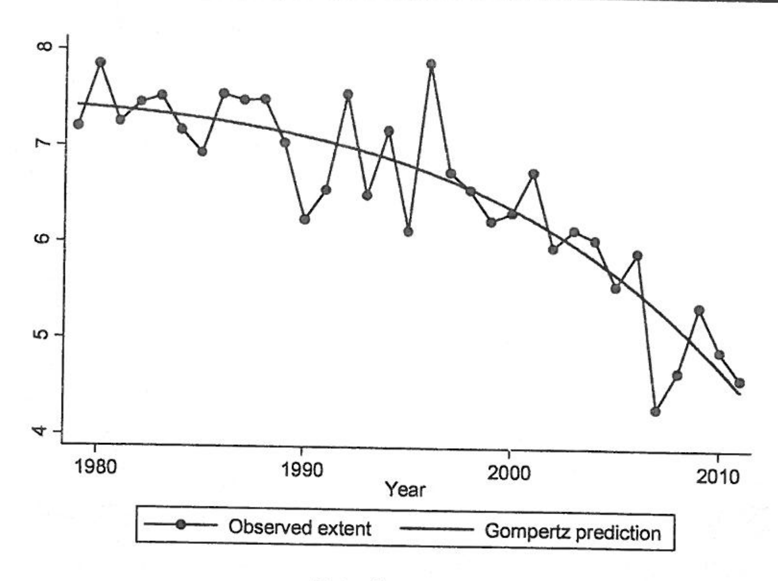
```
(obs = 33)

      Source |       SS       df       MS
-------------+------------------------------         Number of obs =         33
       Model |  1425.43798     3  475.145994         R-squared     =     0.9957
    Residual |  6.15941312    30  .205313771         Adj R-squared =     0.9953
-------------+------------------------------         Root MSE      =   .4531156
       Total |   1431.5974    33  43.3817393         Res. dev.     =   38.25858

3-parameter Gompertz function, extent = b1*exp(-exp(-b2*(year - b3)))
------------------------------------------------------------------------------
      extent |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         /b1 |   7.580278    .291652    25.99   0.000     6.984645    8.175911
         /b2 |  -.0995915   .0271646    -3.67   0.001     -.155069   -.044114
         /b3 |   2017.531   2.173212   928.36   0.000     2013.093    2021.969
------------------------------------------------------------------------------
```

نموذج جومبرتز مناسب بشكل كبير، حيث إن الثلاث معلمات ذات معنوية إحصائية، المعلمة الأولى b1=7.58 تعطي نقطة بداية تقريبية للنموذج وهي 7.58 مليون كم^2. المعلمة الثانية b2=-0.0996 تتحكم في التغير في معدل الانخفاض. المعلمة الثالثة b3=2017.5 تعطي نقطة انقلاب التي عندها ينتقل المنحنى من الارتفاع (معدل انخفاض مرتفع) إلى محدّب لأعلى (معدل انخفاض ضعيف) خلال سنة 2017. لتمثيل هذا النموذج بيانياً الشكل (10.8) يعرض القيم المتوقعة، والتي تم توصيلها بواسطة منحنى وسيط (منحنى التجانس).

```
.predict gomext1
.graph twoway connect extent year
    || mspline gomext1 year, band(50)
    lwidth(medthick)
    legend(label(1 "Observed extent")
    label(2 "Gompertz prediction"))
```



الشكل (10.8)

منحنى جومبرتز في الشكل (10.8) لا يبدو مختلفاً عن المنحنى الربيعي (لم يتم عرضه هنا) ويتناسب بدرجة بسيطة، كما تنقصه زيادة طفيفة وغير واقعية في المنحنى الربيعي خلال السنوات الأولى، وعموماً يبدو أن هناك اختلافات جوهرية تظهر عند استقراء نتائج منحنى جومبرتز خارج نطاق البيانات.

وإذا افترضنا أننا نفكر في إضافة بيانات جديدة في التحليل ليشمل الفترة حتى سنة 2030 – البيانات الواقعية لدينا تشمل حتى سنة 2011 – فإننا نبدأ بإضافة 19 مشاهدة إضافية لاتحتوي على بيانات الجليد، ولكن تحتوي على قيم السنة الجديدة فقط *year* سنة 2012 وحتى سنة 2030. الأمر الأول في الأوامر أدناه يحدد عدد المشاهدات وهو 52 (والذي كان 33)، الأمر الثاني يحسب قيم *year* وهي تساوي السنة الماضية زائداً 1 لكل مشاهدة لاتوجد بها قيمة للسنة، وأخيراً نقوم بإعادة تقدير نموذج جومبرتز مع السنوات الجديدة وسوف نحصل على نفس النتائج التي حصلنا عليها سابقاً.

```
.set obs 52
.replace year = year[_n-1]+1 if year==.
.sort year
.nl gom3: extent year, nolog
```

```
(obs = 33)

      Source |       SS       df       MS
-------------+------------------------------         Number of obs =         33
       Model |  1425.43798     3  475.145994         R-squared     =     0.9957
    Residual |  6.15941312    30  .205313771         Adj R-squared =     0.9953
-------------+------------------------------         Root MSE      =   .4531156
       Total |   1431.5974    33  43.3817393         Res. dev.     =   38.25858

3-parameter Gompertz function, extent = b1*exp(-exp(-b2*(year - b3)))
------------------------------------------------------------------------------
      extent |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         /b1 |   7.580278    .291652    25.99   0.000     6.984645    8.175911
         /b2 |  -.0995915   .0271646    -3.67   0.001     -.155069   -.044114
         /b3 |   2017.531   2.173212   928.36   0.000     2013.093    2021.969
------------------------------------------------------------------------------
```

بالرغم من أن بيانات الجليد، وتركيبة النموذج لم تتغير، فإن البيانات الجديدة امتدت لتشمل سنوات إضافية، مما يسهّل عملية الحصول على قيم متوقعة لكل السنوات من 1979 وحتى 2030.

```
.predict gomext2
```

بخلاف القيم المتوقعة، فإن البواقي للمتغير *extent* يتم حسابها للسنوات التي تحتوي على بيانات فقط.

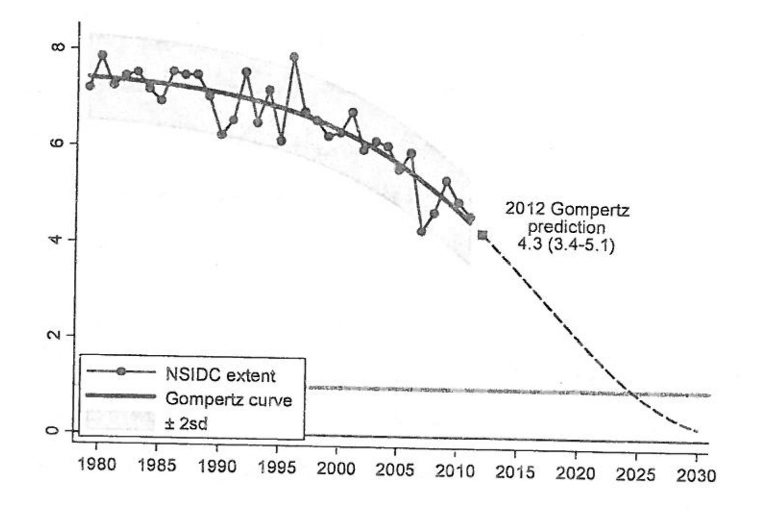
```
.predict res, resid
.summarize res
```

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         res |        33    .0000746    .4387273  -1.039796   1.137713
```

الأوامر أدناه تستخدم القيم المتوقعة الجديدة (*gomext2*) مع الانحراف المعياري للبواقي. الخيار **r(sd)** (تم تعريف النتائج بواسطة الأمر **summarize**) يُستخدم لتحديد الحدود العليا والدنيا لفترات الثقة ±2sd حول القيم المتوقعة. ثم بعد ذلك نقوم بإنشاء رسم بياني يعرض منحنى جومبرتز للفترة الممتدة حتى 2030 مع التأكيد على التوقع الخاص بسنة 2012.

```
.gen gomlo = gomext2 -2*r(sd)
.gen gomhi = gomext2 +2*r(sd)
.label variable gomext2 "nl gom3: extent year"
.label variable gomlo "Gompertz extent - 2sd"
```

```
.label variable gomlo "Gompertz extent + 2sd"
.graph twoway rarea gomlo gomhi year if year<
 2012, color(gs13)
    || mspline gomext2 year if year< 2012,
    bands(60)
    lwidth(thick) lcolor(maroon)
    || mspline gomext2 year if year>= 2012,
    bands(60)
    lwidth(medthick) lcolor(maroon)
    lpattern(dash)
    || connect extent year, lwidth(medthick)
    msymbol(O)
    lcolor(navy) mcolor(navy)
    || scatter gomext2 year if year == 2012,
    msymbol(S)
    mcolor(maroon)
    || if year>1978, xlabel(1980(5)2030, grid)
    yline(0, lcolor(black))
    yline(1, lcolor(gs11) lwidth(thick))
    xtitle("") legend(order(4 2 1) label(2
    "Gompertz curve")
    label(4 "NSIDC extent") label(1
    "`=char(177)' 2sd")
    position(7) ring(0) col(1))
    text(4.5 2019 "2012 Gompertz" "prediction"
    "4.3 (3.4`=char(150)'5.1)", color(maroon))
```



الشكل (11.8)

فترة الثقة التي تظهر باللون الرمادي الفاتح في الشكل (11.8) تم إنشاؤها أولاً، وقام الأمر **twoway rarea** (مدى الفترة) بالتمثيل البياني للسنوات ما قبل 2012، ثم بعد ذلك تم إنشاء منحنى متوسط ذي لون أحمر داكن (**mspline**) يمثل القيم المتوقعة لجومبرتز للفترة ما قبل 2012 وتم وضع هذا المنحنى فوق فترة الثقة الرمادية. بعد ذلك تم وضع قيم المتغير *extent* في الشكل، وتم تمثيلها بخط متصل (**connect**). الخطوة الخامسة والأخيرة تضمنت وضع ثوابت شكل الانتشار، وهي نقطة واحدة تمثل القيم المتوقعة لسنة 2012. أما النصوص الموجودة في الرسم فهي تحدد القيمة الرقمية وحدود فترة الثقة لهذا التوقّع، وكان لون الخط الخاص بالنصوص في الرسم البياني أحمر داكناً وهو نفس اللون لمنحنى القيم المتوقعة، هناك أيضاً خط أفقي رمادي عند نقطة 1 مليون كم^2 (**yline(1)**) تمثل المستوى المنخفض للجليد خلال شهر سبتمبر، والذي يُعتبر فترة خالية من الجليد.

يجب التأكيد على أن منحنيات التنبؤ بهذا النمط لاتمثل طريقة موثوقة لتوقع المستقبل. وهذا النموذج بالذات لا يعطي فهماً مادياً لما سوف يحدث فمثلاً لا يعطي ماهو سبب انخفاض الجليد، فهذا يعتمد بالكامل على اختياراتنا الأوّلية للمتغيرات في النموذج الإحصائي، وكيف تتناسب هذه المتغيرات مع البيانات التاريخية؛ وكتطبيق إحصائي يمكننا تطوير هذه التوقعات البسيطة خطوة أخرى، المنحني في الشكل (11.8) ينخفض إلى أقل من 1 مليون كم^2 في سنة 2025 معطياً نقطة توقع عملية للمناطق الخالية من الجليد في المنطقة القطبية الشمالية، ولكن حتى ولو ظهر أن نموذج جومبرتز إلى حد ما صحيح، فإننا يجب أن نتوقع بأن السلوك الواقعي لهذا المنحنى يتغير كما حصل في الماضي.

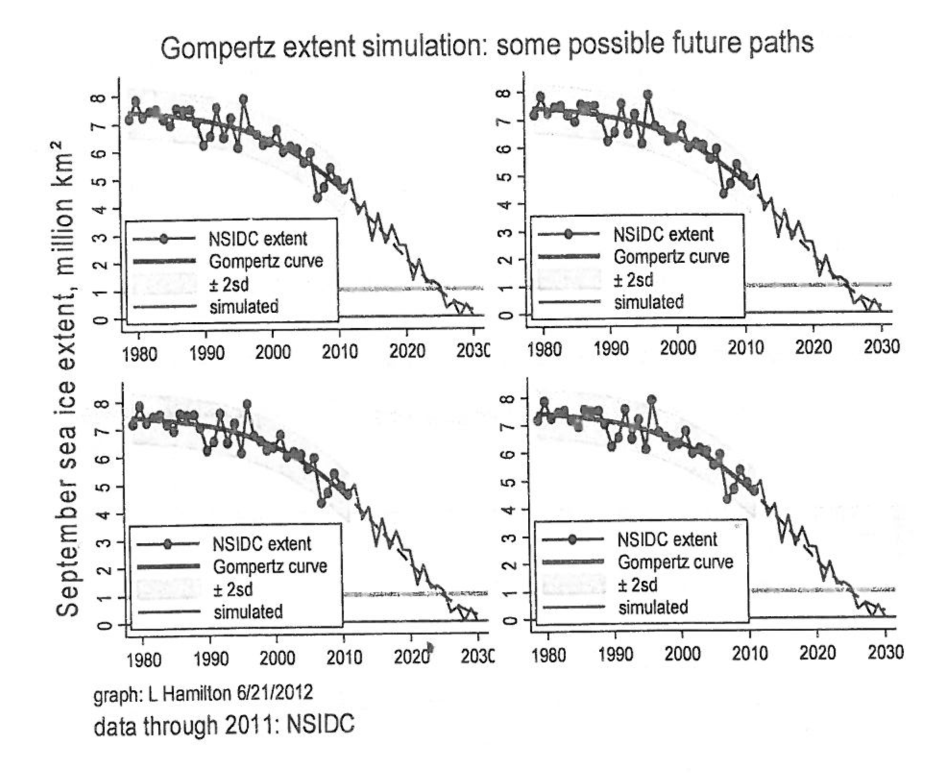
بافتراض أن التباين حول المنحنى للسنوات القادمة يتبع التوزيع الطبيعي مع بعض الانحراف المعياري الذي يظهر في المنحنى في الفترات الماضية فإنه يمكننا محاكاة مثل هذا السلوك بإضافة تذبذب إلى منحنى التجانس واستخراج قيم عشوائية من توزيع طبيعي له انحراف معياري يساوي بواقي الفترات الماضية. فمثلاً الأوامر أدناه تقوم بإنشاء مجموعة جديدة من القيم

المتوقعة (*gomext3*) زائداً تذبذبات طبيعية عشوائية مع الانحراف المعيـــاري للبواقي (**r(sd)**) بعد الأمر *res* **summarize** وإذا كان التنبؤ الجديد يُشير إلـــى مدى سالب وهو مستحيل فعلياً، فإننا نعدل القيمة لتكون صفرًا فقط وليـــست أقل من ذلك.

```
.quietly summ res
.gen gomext3 = gomext2 + r(sd)*rnormal()
.replace gomext3 = 0 if gomext3< 0
.replace gomext3 = extent if year == 2011
```

الأوامر أدناه سوف تقوم بإنشاء رسم بياني للنتائج بطريقة مشابهة للشكل (10.8).

```
.graph twoway rarea gomlo gomhi year if year<
 2012, color(gs13)
     || mspline gomext2 year if year< 2012,
     ands(60)
     lwidth(thick) lcolor(maroon)
     || mspline gomext2 year if year>= 2012,
     bands(60)
     lwidth(medthick) lcolor(maroon)
     lpattern(dash)
     || connect extent year, lwidth(medthick)
     msymbol(O)
     lcolor(navy) mcolor(navy)
     || line gomext3 year if year>= 2011,
     lwidth(medthick)
     lcolor(midblue)
     || , xlabel(1980(10)2030, grid)
     yline(0, lcolor(black)) yline(1,
     lcolor(gs11) lwidth(thick))
     ylabel(0(1)8) ytitle("") xtitle("")
     legend(order(4 2 1 5) label(2 "Gompertz
     curve")
     label(4 "NSIDC extent") label(1
     "`=char(177)' 2sd")
     label(5 "simulated") position(7) ring(0)
     col(1) rowgap(*.3))
```



الشكل (12.8)

في كل مرة نقوم بها بإدخال الأوامر أعلاه، سوف نحصل على قيم متذبذبة عشوائية مختلفة، وبالتالي رسم بياني مختلف. الشكل (12.8) يجمع أربعة أشكال بيانية في شكل واحد يساعد في التأكيد على الانحراف غير المتوقع من سنة لأخرى. امتداد الجليد في أي سنة قد يرتفع أو ينخفض، وحيث إن المنحنى يعرض متوسط التغيرات، فمن غير المتوقع أن تتطابق هذه الأشكال.

أحياناً الرسومات البيانية المنشورة تبقى لفترة طويلة، ويتم تداولها لأهداف لم ينو الكاتب القيام بها، واستخدامها بدون علم الكاتب نفسه. لذلك فإنه من الأفضل أن تقوم بوضع اسمك، ومصدر البيانات، وأي معلومات إيضاحية أخرى على الرسم البياني نفسه، كما هو واضح في الشكل (12.8)

باستخدام الخيارات note و caption، الأمر أدناه يقوم بدمج أربعة أشكال في صورة واحدة تم تسميتها *Gompertz_extent1* و *Gompertz_extent2* وهكذا.

```
.graph combine Gompertz_extent1.gph
 Gompertz_extent2.gph
     Gompertz_extent3.gph Gompertz_extent4.gph,
     title("Gompertz extent simulation: some
    possible future
     paths", size(medlarge))
     caption("data through 2011: NSIDC")
     note("graph: L Hamilton 6/21/2012")
    imargin(small) col(2)
     l1title("September sea ice extent, million
    km`=char(178)'")
```

انحدار بوكس-كوكس : Box-Cox Regression

نترك بيانات المناطق الباردة خلفنا، ونبدأ بالعمل على بيانات التنمية البشرية للأمم المتحدة في بقية هذا الفصل. هذه البيانات موجودة بالملف *Nations3.dta*، العلاقات غير الخطية بين المتغيرات واضحة في أشكال الانتشار لهذه البيانات، وهذه الأشكال مثل الشكل (4.7)، اللوغاريتمات وطرق التحويل الأخرى من سلّم توكي للقوى Tukey's ladder (تم الإشارة إليه في الفصل 5) تُعتبر أدوات بسيطة تجعل العلاقات غير الخطية أكثر خطية، مما يُمكّننا من تطبيق OLS والانحدار الموثوق والنماذج الخطية الأخرى.

اختيار استخدام أي طريقة تحويل قد يتضمن محاولة عدة خيارات واختبار كيف. إن كل خيار يؤثر في التوزيعات وأشكال الانتشار والبواقي. هناك مدخل منهجي يُسمى انحدار بوكس - كوكس بدلاً من استخدام تقدير الأرجحية العظمى لاختيار معلمات تحويل بوكس - كوكس التي تُعتبر أفضل

نموذج انحدار معين. تحويل بوكس – كوكس الأكثر استخداماً يتم تطبيقه على كل المتغيرات في النموذج أو أي مجموعة فرعية من المتغيرات.

ولنأخذ على سبيل المثال انحدار متوسط العمر المتوقع على ستة متغيرات أخرى تتضمن *reg1*، ومتغير وهمي يتم إنشاؤه عن طريق الأمر **tabulate** والذي يُشير إلى الدول الأفريقية.

```
.use C:\data\Nations3.dta, clear
.describe life adfert urban gdp chldmort school
 reg1
```

variable name	storage type	display format	value label	variable label
life	float	%9.0g		Life expectancy at birth 2005/2010
adfert	float	%8.0g		Adolescent fertility: births/1000 fem 15-19, 2010
urban	float	%9.0g		Percent population urban 2005/2010
gdp	float	%9.0g		Gross domestic product per cap 2005$, 2006/2009
chldmort	float	%9.0g		Prob dying before age 5/1000 live births 2005/2009
school	float	%9.0g		Mean years schooling (adults) 2005/2010
reg1	byte	%8.0g	reg1	region==Africa

ليست هناك فائدة في تحويل متغير ثنائي القيم مثل *reg1* ولكن شكل الانتشار للمتغيرات الأخرى يعرض أنماطاً غير خطية تجعلها مرشحة للتحويل. في هذا المثال، نقوم بتحليل انحدار المتغير *life* على المتغيرات *adfert, urban, gdp, chldmort, school, reg1*، المتغير المستقل أو متغير الطرف الأيسر هو *life* والذي يكون في العادة في شكل صف، ولكن التحويل مطلوب لكل المتغيرات في الطرف الأيمن باستثناء المتغير *reg1*. هذه الاختيارات يمكن القيام بها باستخدام الخيارات **model(rhsonly)** **notrans(*reg1*)**، ويجب ملاحظة أن قائمة المتغيرات الأولية لا تتضمن المتغير *reg1* حيث تم تحديده بشكل منفصل في الخيار **notrans()**

```
.boxcox life adfert urban gdp chldmort school,
 model(rhsonly) notrans(reg1)
```

```
Fitting full model

Iteration 0:   log likelihood = -455.39883  (not concave)
Iteration 1:   log likelihood = -430.26519
Iteration 2:   log likelihood = -429.92904
Iteration 3:   log likelihood = -429.92798
Iteration 4:   log likelihood = -429.92798
(15 missing values generated)
(1 missing value generated)
(6 missing values generated)
(15 missing values generated)
(1 missing value generated)
(6 missing values generated)
(15 missing values generated)
(1 missing value generated)
(6 missing values generated)

                                                  Number of obs   =        178
                                                  LR chi2(7)      =     463.38
Log likelihood = -429.92798                       Prob > chi2     =      0.000
```

life	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
/lambda	.4867359	.0513717	9.47	0.000	.3860492	.5874226

Estimates of scale-variant parameters

	Coef.
Notrans	
reg1	-2.863907
_cons	86.17721
Trans	
adfert	-.0383667
urban	.2065436
gdp	.000283
chldmort	-1.42784
school	-1.601755
/sigma	2.708479

Test H0:	Restricted log likelihood	LR statistic chi2	P-value Prob > chi2
lambda = -1	-524.20312	188.55	0.000
lambda = 0	-476.81642	93.78	0.000
lambda = 1	-455.39883	50.94	0.000

قيمة (لمدا) المعطاة في جدول المخرجات أعلاه (λ = 0.4867359) هـي المعلمة المختارة للشكل العام لتحويلات بوكس - كوكس

$$X^{(\lambda)} = \{x^{\lambda}-1\}/\lambda$$

مُعَامِلات الانحدار المعروضة في جدول مخرجات بوكس - كوكس أعلاه تتعلق بانحدار عادي لمتغيرات تم تحويلها في هذا النمط، بالإمكان تكرار نتائج انحدار بوكس - كوكس بواسطة إنشاء نسخة محوّلة من كل متغير ثم استخدام الأمر **regress**.

```
.gen bc2adf = (adfert^.4867359-1)/.4867359
.gen bc2urb = (urban^.4867359-1)/.4867359
.gen bc2school = (school^.4867359-1)/4867359
.gen bc2gdp = (gdp^.4867359-1)/.4867359
.gen bc2chld = (chldmort^.4867359-1)/.4867359
.regress life bc2* reg1
```

Source	SS	df	MS
Model	16332.407	6	2722.06783
Residual	1305.78282	171	7.63615683
Total	17638.1898	177	99.6507898

Number of obs =	178
F(6, 171) =	356.47
Prob > F =	0.0000
R-squared =	0.9260
Adj R-squared =	0.9234
Root MSE =	2.7634

life	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
bc2adf	-.0383667	.0598489	-0.64	0.522	-.1565045	.0797711
bc2urb	.2065436	.0955322	2.16	0.032	.0179693	.395118
bc2school	-1.60e+07	3209717	-4.99	0.000	-2.24e+07	-9681777
bc2gdp	.000283	.0035913	0.08	0.937	-.006806	.0073721
bc2chld	-1.42784	.0792446	-18.02	0.000	-1.584263	-1.271416
reg1	-2.863907	.674814	-4.24	0.000	-4.195945	-1.531869
_cons	86.17721	1.909628	45.13	0.000	82.40773	89.94669

انحدار بوكس - كوكس وجد أن معلمة التحويل λ هي الأمثل في سياق معيار الأرجحية العظمى. مقابلة متطلبات هذا المعيار لا تعني بالضرورة أن العلاقات أصبحت خطية. الهدف الأخير - وهو جعل العلاقات خطية - ربما من الأفضل استمرار محاولة القيام به من خلال استخدام الفحص المرئي والحكم الشخصي مع احتمالية استخدام تحويلات مختلفة للمتغيرات.

الإسناد المتعدد للقيم المفقودة :
Multiple Imputation of Missing Values

ملف البيانات *Nations3.dta* يحتوي على معلومات عن 194 دولة، ولكن القيم المفقودة تقيّد التحليل الذي قمنا به في الجزء السابق ليكون التحليل لمجموعة فرعية مكوّنة من 178 دولة لها معلومات متكاملة لكل المتغيرات. مدخل قائمة الحذف الذكية لاستبعاد القيم المفقودة وهي ممارسة إحصائية شائعة بدافع الضرورة ومن عيوبها المعروفة خسارة بعض المشاهدات، ونقص القوة الإحصائية، وإذا كانت المشاهدات ذات القيم المفقودة تختلف بدرجة كبيرة عن باقي المشاهدات، فإن قائمة الحذف الذكية قد تؤدي إلى تحيّز المُعَامِلات المقدّرة.

قد تكون هناك متغيرات أخرى في البيانات التي ترتبط إحصائياً بالقيم المفقودة. في مثل هذه الحالات، فإن الانحدار يمكن أن يُستخدم لتوقّع ما هي القيم المفقودة وهذه التوقعات تُستخدم كبديل للقيم المفقودة في خطوات تحليلية تالية. إسناد الانحدار للقيم المفقودة يمكنه استعادة المشاهدات والقوة الإحصائية الظاهرية، وتقليل احتمالية الحصول على مُعَامِلات متحيّزة. وعموماً فإن القيم المُسندة سوف يكون لها تباين منخفض عن تلك القيم الموجودة لأي متغير، مما يؤدي إلى تقديرات خطأ معياري متحيزة تقترب من الصفر. بعبارة أخرى إسناد الانحدار قد يتسبب في الحصول على تقديرات مبالغ فيها في الدقة أو المعنوية الإحصائية للنتائج المحسوبة.

الإسناد المتعدد للقيم المفقودة يبدأ من الفكرة الأساسية لإسناد الانحدار ثم تتم إضافة خطوات أخرى للحصول على تقديرات واقعية للأخطاء المعيارية أو عدم التأكد. هذه الخطوات تتضمن إنشاء مجموعات متعددة من المشاهدات الوهمية تُستخدم كبديل للقيم المفقودة، وذلك عن طريق توقعات الانحدار زائداً تذبذبات عشوائية. ثم الخطوة الأخيرة هي جمع معلومات الإسناد المتعدد لتقدير نموذج الانحدار مع أخطائه المعيارية واختباراته.

مجموعة أوامر ستاتا **mi** لإجراءات الإسناد المتعدد تدعم عددًا من طرق تنظيم البيانات، وطرق التقدير، وتقنيات إنشاء النماذج بما فيها النماذج

اللوغاريتمية للمتغيرات النوعية، دليل المُـستخدم *Stata Multiple-Imputation Reference Manual* يغطي هذا الخيارات، بالإضافة إلى عدد آخـر إضـافي يحتوي على أمثلة أكثر تعقيداً.

وكمثال بسيط، يمكننا العودة إلى تحليل انحدار متوسط العمر المتوقع.

```
.use C:\data\Nations3.dta, clear
.regress life adfert urban loggdp chldmort
 school reg1
```

Source	SS	df	MS
Model	15810.6966	6	2635.1161
Residual	1827.49317	171	10.6870946
Total	17638.1898	177	99.6507898

Number of obs =	178
F(6, 171) =	246.57
Prob > F =	0.0000
R-squared =	0.8964
Adj R-squared =	0.8928
Root MSE =	3.2691

life	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
adfert	-.0039441	.0091498	-0.43	0.667	-.0220053	.014117
urban	.0397506	.016287	2.44	0.016	.0076012	.0718999
loggdp	2.90728	.9196223	3.16	0.002	1.092007	4.722554
chldmort	-.1314439	.0102063	-12.88	0.000	-.1515905	-.1112972
school	-.3322321	.1480558	-2.24	0.026	-.6244844	-.0399798
reg1	-3.56938	.7845902	-4.55	0.000	-5.118109	-2.02065
_cons	65.3779	3.124978	20.92	0.000	59.2094	71.5464

هناك ثلاثة متغيرات في هذا التحليل – *loggdp, chldmort, school* – بها قيم مفقودة، وهذه المتغيرات مجتمعة تؤدي إلى انخفاض مشاهدات العينة من 194 إلى 178 مشاهدة.

```
.summarize life adfert urban loggdp chldmort
 school reg1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
life	194	68.7293	10.0554	45.85	82.76666
adfert	194	51.81443	44.06612	1	207.1
urban	194	55.43488	23.4391	10.25	100
loggdp	179	3.775729	.5632902	2.446848	4.874516
chldmort	193	47.65026	52.8094	2.25	209
school	188	7.45922	2.959589	1.15	12.7
reg1	194	.2680412	.4440852	0	1

الأمر **misstable summarize** يحسب ثلاثة أنواع من المشاهدات بناءً على حالة القيم المفقودة لتلك المشاهدات:

obs=. القيمة المفقودة الافتراضية لبرنامج ستاتا، ويُشار إليها بـ "المفقود الناعم".

obs>. رموز القيم المفقودة تُعرض كحروف مع نقاط مثل a. ,b. ,c. الخ، ويمكن إضافة توصيف لها، ويُشار إليها بـ "المفقود الخشن".

obs<. قيم موجودة.

يمكن لبرنامج ستاتا إدخال قيم مفقودة ناعمة فقط، ولا يمكنه إدخال القيم الخشنة. وتجب الإشارة إلى أن جميع القيم المفقودة في ملف البيانات *Nations3.dta* هي من النوع الناعم، ولذلك فإن الوضع أسهل. وهناك مثال لدراسة استقصائية تحتوي على قيم خشنة سوف يتم تناوله في الفصل (9).

```
.misstable summarize life adfert urban loggdp
 chldmort school reg1
```

```
(UN Human Development Indicators)

. misstable summarize life adfert urban loggdp chldmort school reg1
                                                                Obs<.
```

Variable	Obs=.	Obs>.	Obs<.	Unique values	Min	Max
loggdp	15		179	179	2.446848	4.874516
chldmort	1		193	144	2.25	209
school	6		188	165	1.15	12.7

الخطوة الأولى في الإسناد المتعدد هي تحديد البيانات باستخدام الأمر **mi set**، والذي يحدد كيف يتم تنظيم القيم المدخلة. هناك أربع طرق محتملة – تم شرحها في دليل المستخدم *Reference Manual*– ففي مثالنا الحالي سوف نختار طريقة الذاكرة الفعّالة **mlong**، حيث إن المشاهدات الجديدة أو الصفوف سوف تُضاف للبيانات.

```
.mi set mlong
```

عند تسجيل الإسناد المتعدد، فإن القيم المفقودة في البيانات الأصلية غير المُسندة سوف تُسجل على أنها m=0، عمليات الإسناد مع مجموعات من القيم التي تم إدخالها كقيم مفقودة تُسجل على أنها $m=3, m=2, m=1$ وهكذا، حرف M يُشير إلى عدد عمليات الإسناد التي تم القيام بها. وقبل التقدم أكثر نحن نحتاج إلى تسجيل المتغيرات التي نريد إسنادها في واحد من الأنواع الثلاثة التالية:

imputed المتغير له قيم مفقودة تحتاج إلى إسناد.

passive متغير وهو عبارة عن دالة للمتغيرات المُسندة أو دالة لمتغيرات سلبية، وهذا المتغير سوف تكون له قيم مفقودة في البيانات الأصلية (m=0) وقيم متفاوتة لكل إسناد ($m=1$، $m=2$ وهكذا).

regular لا المتغيرات المسندة ولا السلبية تكون لها نفس القيم، (المفقودة أو الموجودة) لكل m.

مثالنا الحالي له قيم مفقودة بالمتغيرات *loggdp, chldmort, school*، لذلك تم تسجيل هذه المتغيرات كمتغيرات مسندة **imputed**

```
.mi register imputed loggdp chldmort school
```

المتغيرات الأخرى *life, adfert, urban, reg1* لا تحتوي على قيم مفقودة، ويجب أن تُسجّل كمتغيرات منتظمة **regular**.

```
.mi register regular life adfert urban reg1
```

الخطوة التالية تقوم بالإسناد الفعلي، والقيم المفقودة للمتغيرات *loggdp, chldmort, school* (متغيرات**mi register imputed**) يتم إسنادها عن طريق انحدار متغيرات **mi register regular** وهي *reg1, adfert, urban*، سوف نستخدم طريقة الانحدار الطبيعي متعدد المتغيرات (**mvn**) وسوف يكون هناك 50 إسنادًا مستقلاً يُشار إليها كـ m=0 (وهي البيانات الأصلية التي تحتوي على قيم مفقودة) أو m=1 وحتى m=50 وكل منها تحتوي على إسنادات لـ 16 مشاهدة كانت أصلاً قيماً مفقودة في البيانات الأصلية. لذا فإن 50×16=800 مشاهدة سوف تُضاف إلى البيانات، مما يجعل مجموع المشاهدات 194+800=964 مشاهدة.

```
.mi impute mvn loggdp chldmort school = adfert
 urban reg1, add(50) rseed(12345)

Performing EM optimization:
  observed log likelihood = -780.80745 at iteration 6

Performing MCMC data augmentation ...

Multivariate imputation                    Imputations =      50
Multivariate normal regression                   added =      50
Imputed: m=1 through m=50                      updated =       0

Prior: uniform                              Iterations =    5000
                                               burn-in =     100
                                               between =     100
```

Variable	Observations per m			
	Complete	Incomplete	Imputed	Total
loggdp	179	15	15	194
chldmort	193	1	1	194
school	188	6	6	194

```
(complete + incomplete = total; imputed is the minimum across m
 of the number of filled-in observations.)
```

الحصول على 50 قيمة مُسندة مستقلة كل منها تحتوي على قيم مفقـــودة تم استبدالها، هذا يوفر قاعدة للتقديرات التي سوف نقوم بها لاحقاً للاختلافات من عينة لأخرى عند قيامنا بجمع هذه القيم للانحدار. الخيــار **rseed(12345)** في الأمر **mi impute** يقوم بتحديد نظام اختياري لمولّد الأرقــام العــشوائية ببرنامج ستاتا. يمكننا إنشاء مثال قابل للتكرار عن طريق اســتخدام الخيــار **rseed()**، وهذا المثال قد يكون مقبولاً لأهداف تعليمية. غير ذلك، فإن ســتاتا سوف يختار الضبط الخاص به مسبباً اختلافات بسيطة فـــي النتـــائج عنـــد استخدام الأمر في المرة التالية.

الخطوة الأخيرة تستخدم هذه الإسنادات لتحليل انحدار متغيــر متوســط العمر المتوقع على 6 متغيرات تنبؤية. مبدئياً فإن عملية الإسناد تؤدي إلـــى تقديرات أكثر كفاءة (أخطاء معيارية منخفضة) وتقلل من التحيّز عند تحليـــل الانحدار الذي استبعد كل المشاهدات التي تحتوي على قيم مفقودة.

.mi estimate, dots: regress *life adfert urban loggdp chldmort school reg1*

```
Imputations (50):
  .........10.........20.........30.........40.........50 done

Multiple-imputation estimates                   Imputations       =         50
Linear regression                               Number of obs     =        194
                                                Average RVI       =     0.0365
                                                Largest FMI       =     0.1167
                                                Complete DF       =        187
DF adjustment:   Small sample                   DF:     min       =     156.85
                                                        avg       =     172.10
                                                        max       =     184.02
Model F test:       Equal FMI                   F(   6,   184.7)  =     240.19
Within VCE type:          OLS                   Prob > F          =     0.0000
```

life	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
adfert	-.0044855	.0091363	-0.49	0.624	-.0225124	.0135414
urban	.047157	.0163199	2.89	0.004	.014945	.079369
loggdp	2.704804	.9755022	2.77	0.006	.7779876	4.63162
chldmort	-.1305611	.0102536	-12.73	0.000	-.1507939	-.1103283
school	-.3317234	.1527427	-2.17	0.031	-.6332567	-.0301902
reg1	-3.814368	.8011135	-4.76	0.000	-5.394917	-2.23382
_cons	65.78153	3.297027	19.95	0.000	59.27041	72.29265

نتائج الأمر **mi estimate** تشبه بدرجة كبيرة تلك التي حصلنا عليها من تحليل الانحدار العادي، هذا يمثل أفضل سيناريو تتشابه فيه الطرق المعقدة والبسيطة، حيث إن النتائج تظهر مستقرة بشكل معقول. وفي أي تقرير بحثي يمكننا عرض التحليل مع الملاحظات التوضيحية لتوضيح أننا استخدمنا مدخلاً آخر للاختبار، وحصلنا على نفس النتيجة.

الفصل (9) يوضح مثالاً ثانياً للإسناد المتعدد باستخدام بيانات دراسة استقصائية بدلاً من نماذج الانحدار الخطية. وللحصول على مزيد من التفاصيل عن هذا الموضوع قم بطباعة الأمر **help mi** أو الاطلاع على دليل المستخدم *Stata Multiple-Imputation Reference Manual*.

نماذج المعادلة الهيكلية : Structural Equation Modeling

نماذج الانحدار السابقة تعاملت مع خصوبة المراهقين، والنسبة المئوية للمناطق الحضرية، والخصائص المحلية الأخرى كمتغيرات تنبؤية محتملة

لمتغير متوسط العمر المتوقع بدون ضرورة التأكيد على أن هذه المتغيـرات هي مسببات لمتوسط العمر المتوقع. أحد هذه المتغيرات التنبؤية هو متغيـر معدل وفيات الأطفال والذي كانت له علاقة سببية بمتوسط العمر المتوقـع، ولكن معدلات وفيات الأطفال بدورها خاضعة للتأثر بالخـصائص المحليـة الأخرى مثل ماهو العمر المتوقع مما يجعل العلاقة السببية أكثر تعقيداً. فعلى سبيل المثال، إذا كانت معدلات الخصوبة لدى المراهقين تؤثر على وفيـات الأطفال ووفيات الأطفال تؤثر على متوسط العمر المتوقع، إذن فإن خصوبة المراهقين لها تأثير غير مباشر على متوسط العمر المتوقـع ومـا إذا كـان متوسط المتوقع له تأثير مباشر واضح. وفي سياق العلاقة السببية، فإن معدل وفيات الأطفال يعمل كمتغير وسيط أو دخيل.

نماذج المعادلة الهيكلية تعتبر طريقة منتظمة لتحليل مثل هذه التـأثيرات غير المباشرة مع الأنواع الأخرى للعلاقات السببية، وتوضيح هذه النمـاذج يظهر في الرسم البياني للمسار الذي يعرض بعض الأفكار حـول الترتيـب والعلاقات السببية. الترتيب السببي للمتغيرات أمر مهم جداً. ونموذج المعادلة الهيكلية لايمكنه إثبات السببية، ولكن يفترض هيكلاً سببياً معيناً ثم يُطبّق عليه التقنيات الإحصائية لعرض التفاصيل وتطوير المحددات بطريقة ما. يجـب علينا الاعتماد على المعرفة الخارجية أو النظرية لتحديد الترتيـب الـسببي البسيط. وإذا كانت المعرفة ضعيفة، فإن التحليل الذي يتبعه سـوف يكـون ضعيفاً أيضاً، ولكن إنشاء رسم بياني للمسار يعتبر خطوة مفيدة حتـى ولـو كان الترتيب غير مؤكد. وفي العادة فإن الرسم البياني يساعد فـي توضـيح الأفكار الغامضة أو شرح أفكارنا.

المداخل التي تعتمد على نماذج المعادلة الهيكلية أصبحت تُهيمن علـى العديد من مجالات مختلفة في البحوث العلمية، نماذج المعادلة الهيكلية تتعلق بشكل خاص بالعلوم الاجتماعية لأنها تريد سد الفراغ الموجود بين المجالات النظرية والبيانات. النماذج أصبحت العنوان الرئيس للكثير من الكتب التـي تناولت قضايا مثل التقدير، ونماذج القياس، وهياكل الخطأ والأسباب المتبادلة (انظر Kline 2010، Skrondal and Rabe-Hesketh 2004)، مع الإصـدار 12

لبرنامج ستاتا قام البرنامج بإضافة أوامر نماذج المعادلة الهيكلية الخاص به، حيث ورد في دليل المستخدم نماذج المعادلة الهيكلية *Structural Equation Modeling Reference Manual*:

نماذج المعادلة الهيكلية ليست فقط طرق تقدير لنموذج معين بنفس طريقة الأمر **regress** *والأمر* **probit** *أو حتى بنفس طريقة الأمر* **stcox** *والأمر* **xtmixed** *ولكن نماذج المعادلة الهيكلية هي طريقة تفكير وكتابة وتقدير.*

هذا الجزء من الكتاب يشرح أمر نماذج المعادلة الهيكلية لبرنامج ستاتا **sem** من خلال توسيع مبسط لانحدار متوسط العمر المتوقع. كنا قد رأينا سابقاً أن متوسط العمر المتوقع تم تقديره بواسطة الخصائص المحلية الأخرى؛ في جدول الانحدار أدناه إحصائيات t وأوزان بيتا أو مُعَاملات الانحدار المعيارية (العمود الموجود في الجانب الأيمن) توضح أن معدل وفيات الأطفال له أكبر تأثير.

```
.use C:\data\Nations3.dta, clear
.regress life adfert urban loggdp chldmort
 school reg1, beta
```

Source	SS	df	MS
Model	88961.8088	6	14826.9681
Residual	16569.5108	971	17.0643778
Total	105531.32	977	108.01568

Number of obs	=	978
F(6, 971)	=	868.88
Prob > F	=	0.0000
R-squared	=	0.8430
Adj R-squared	=	0.8420
Root MSE	=	4.1309

life	Coef.	Std. Err.	t	P>|t|	Beta
adfert	-.0186354	.0061076	-3.05	0.002	-.0542246
urban	.0926001	.0077519	11.95	0.000	.2355879
loggdp	.8178748	.4963185	1.65	0.100	.0425882
chldmort	-.1065984	.0063543	-16.78	0.000	-.4893582
school	-.1540576	.0804122	-1.92	0.056	-.0408058
reg1	-7.910748	.6323653	-12.51	0.000	-.2744216
_cons	68.87724	1.743847	39.50	0.000	.

عند كتابة الأمر **sem** يمكننا إدخال نفس النموذج وذلك كما يلي:

```
.sem (life<- adfert urban loggdp chldmort
 school reg1), standard
```

```
(16 observations with missing values excluded;
 specify option 'method(mlmv)' to use all observations)

Endogenous variables

Observed:  life

Exogenous variables

Observed:  adfert urban loggdp chldmort school reg1

Fitting target model:

Iteration 0:   log likelihood = -18660.532
Iteration 1:   log likelihood = -18660.532

Structural equation model                      Number of obs      =       978
Estimation method  = ml
Log likelihood     = -18660.532
```

Standardized	Coef.	OIM Std. Err.	z	P>|z|	[95% Conf.	Interval]
Structural						
life <-						
adfert	-.0542246	.0176934	-3.06	0.002	-.088903	-.0195462
urban	.2355879	.0194051	12.14	0.000	.1975547	.2736211
loggdp	.0425882	.0257454	1.65	0.098	-.0078719	.0930483
chldmort	-.4893582	.0283438	-17.27	0.000	-.5449109	-.4338055
school	-.0408058	.0212159	-1.92	0.054	-.0823882	.0007767
reg1	-.2744216	.0215576	-12.73	0.000	-.3166737	-.2321695
_cons	6.630625	.1869022	35.48	0.000	6.264304	6.996947
Variance						
e.life	.1570104	.0070122			.1438511	.1713735

```
LR test of model vs. saturated: chi2(0)   =      0.00, Prob > chi2 =      .
```

المُعاملات المعيارية في مخرجات الأمر **sem** تساوي أوزان (بيتا) التي تم الحصول عليها باستخدام الأمر **regress**، وبخلاف الأمر **regress**، فإن الأمر **sem** يعرض الأخطاء المعيارية للمُعاملات المعيارية، وهذا يؤدي إلى الحصول على إحصائيات z تشبه إحصائيات t التي أظهرها الأمر **regress** ولكن مع احتمالات مختلفة اختلافاً بسيطاً بسبب مقارنتها مع التوزيعات الطبيعية المعتدلة بدلاً من توزيعات t، أن كتابة الأمر **sem** بين الأقواس

(*life<- adfert urban loggdp chldmort school reg1*) يحدد المسارات السببية للمتغير *life* مع المتغيرات *adfert, urban* ... الخ.

معدل وفيات الأطفال هو أقوى متغير تنبؤي لمتغير متوسط العمر المتوقع ويمكن تقديره – متوسط العمر المتوقع – من العديد من الخصائص المحلية، أما خصوبة المراهقين والتي أظهرت نتائج لم تكن ذات معنوية إحصائية مع متوسط العمر المتوقع في نتائج الأمر **regress** أو نتائج الأمر **sem** في الجدول أعلاه هي أقوى متغير تنبؤي لمعدل وفيات الأطفال.

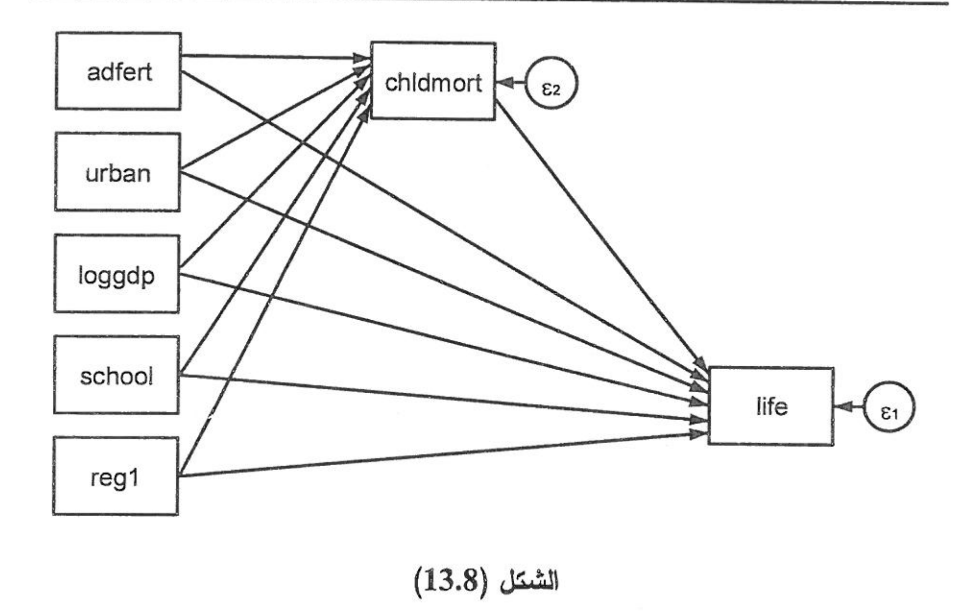
```
.regress chldmort adfert urban loggdp school
 reg1, beta
```

Source	SS	df	MS
Model	1801371.17	5	360274.235
Residual	422622.667	972	434.796983
Total	2223993.84	977	2276.34989

Number of obs =	978
F(5, 972) =	828.60
Prob > F =	0.0000
R-squared =	0.8100
Adj R-squared =	0.8090
Root MSE =	20.852

chldmort	Coef.	Std. Err.	t	P>|t|	Beta
adfert	.2318316	.0299192	7.75	0.000	.1469444
urban	.0245631	.0391215	0.63	0.530	.0136128
loggdp	-15.56996	2.455012	-6.34	0.000	-.1766092
school	-5.479663	.3658743	-14.98	0.000	-.316167
reg1	57.32674	2.609213	21.97	0.000	.4331934
_cons	128.5363	7.777316	16.53	0.000	.

الشكل (13.8) يعرض رسماً بيانياً للمسار الذي يظهر به متغير معدل وفيات الأطفال كمتغير دخيل تأثر بمتغير خصوبة المراهقين والخصائص الأخرى، وهو أيضاً متغير تنبؤي لمتوسط العمر المتوقع. ومن الناحية النظرية، فإن السبب يتجه من اليسار إلى اليمين في الشكل أدناه، والتأثيرات غير المباشرة يمكن أن تتبع أي مسارات تتراوح من المتغيرات التي تمثل الخصائص المحلية للمجتمع إلى متغير معدل الوفيات *chldmort* ومن متغير *chldmort* إلى متغير *life*، المربعات الظاهرة في الرسم أدناه تمثل متغيرات منظورة في النموذج، سوف يتم التطرق إلى موضوع المتغيرات غير المنظورة أو المخفية في الفصل (12).



الشكل (13.8)

الشكل (13.8) أعلاه تم إنشاؤه باستخدام واجهة المستخدم البيانية graphical user interface(GUI) والتي تسمى SEM Builder، ويمكن الحصول على مربع حوار واجهة المستخدم البيانية من خلال طباعة الأمر:

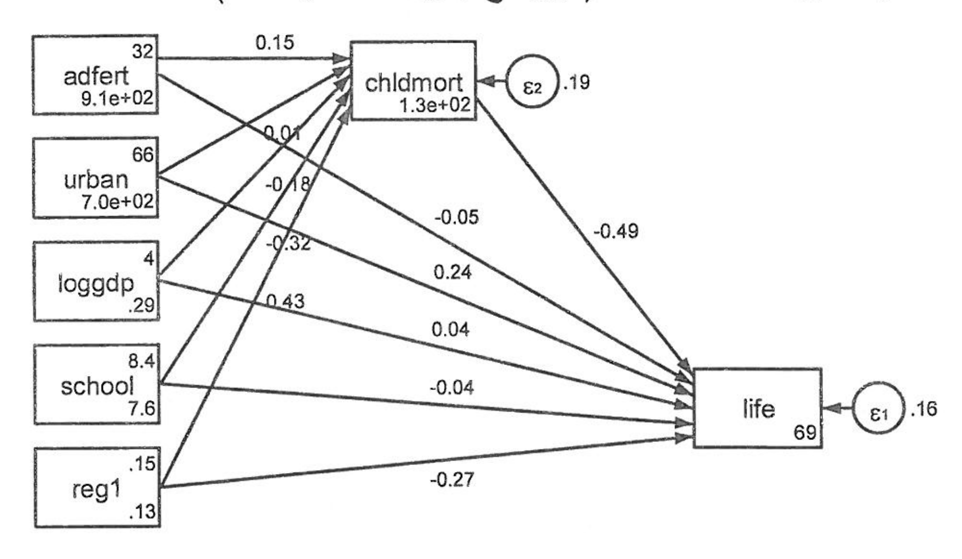
```
.sembuilder
```

أو بواسطة استخدام قائمة ستاتا:

```
Statistics > SEM (structural equation modeling)
> Model building and estimation
```

للحصول على معلومات عن كيفية البدء في استخدام واجهة المستخدم البيانية قم بطباعة الأمر **help sembuilder**. العناصر الرئيسة في الشكل (13.8) هي المتغيرات المنظورة، أول شيء لإنشاء الرسم البياني تتم بإضافة مربعات خالية باستخدام أداة *إضافة متغير ملحوظ* Add Observed Variable من القائمة الموجودة في يسار الشاشة، ثم بعد ذلك تتم إضافة أسماء المتغيرات عن طريق اختيار كل مربع باستخدام أداة Select، وإضافة الأسماء من خلال القائمة المنسدلة Variable، أداة *إضافة المسار* Add Path يمكنها ربط المتغيرات، وما عليك إلا النقر على المتغير الخارجي أو المسبب ثم سحبه إلى السهم ليربط المتغير الداخلي أو المتأثر.

من قائمة شريط الأدوات الموجودة في أعلى الإطار قم باختيار Estimation > Estimate للحصول على المُعامِلات والإحصائيات الأخرى للنموذج. الشكل (14.8) يعرض النتائج الافتراضية لتقدير الشكل (14.8)، وكل مسار متعلق مع مسار معياري أو معامل انحدار مثل معامل انحدار *adfert* على *chldmort* 0.23 (قارن مع جدول الانحدار أعلاه).

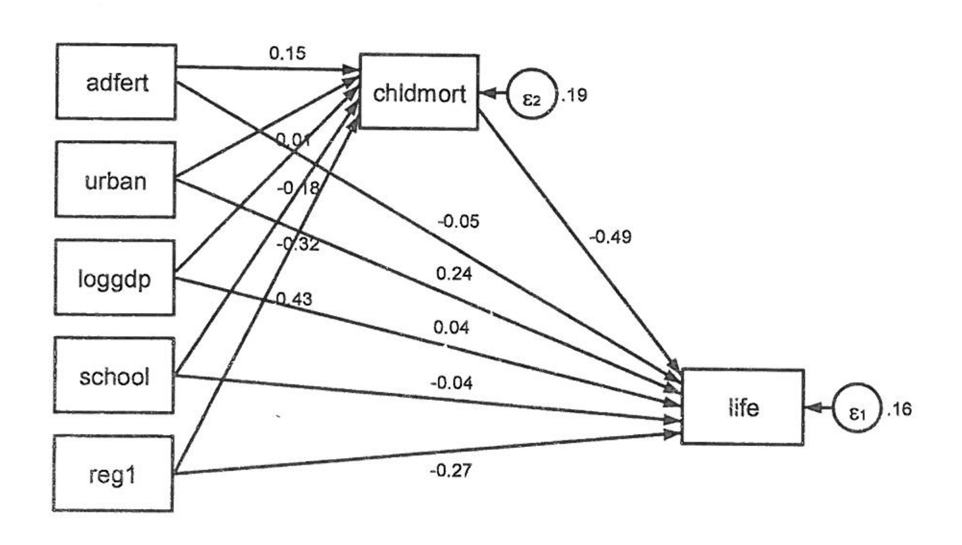


الشكل (14.8)

في كل مربع لكل متغير خارجي في الشكل (14.8) لدينا متوسط وتباين للمتغير. التحليل الموجود لدينا يحتوي على 178 دولة، متوسط المتغير *adfert* يساوي 32 تقريباً، والتباين يساوي 9.1e+02=910، مربعات المتغيرات الخارجية تحتوي على قيم تقاطعها مع المحور العمودي *y* مثل 1.3e+02=130 للمتغير *chldmort*. مرة أخرى قم بالمقارنة مع نتائج الانحدار السابق، وأخيراً الشكل (14.8) يعطي تباين البواقي المتعلقة بالخطأ العشوائي ε1(*chldmort*) و ε2(*life*).

وهناك إصدار أكثر بساطة للشكل أعلاه يحتوي على مُعَامِلات المسار المعياري، وتباين البواقي المعيارية في الشكل (15.8)، هذا التبسيط يمكن القيام به عن طريق قائمة الاختيارات الموجودة في أعلى إطار SEM Builder

```
Settings > Variables > All … > Results >
Exogenous variables > None > OK
Settings > Variables > All … > Results >
Endogenous variables > None > OK
Settings > Variables > Error … > Results >
Error std. variance > OK
Settings > Connections > Paths > Results > Std.
parameter > OK
Settings > Connections > All > Results > Result
1 > Format %3.2f > OK > OK
Settings > Connections > All > Results > Result
2>None > OK
Estimation > Estimate > OK
```



الشكل (15.8)

الشكل (15.8) يتعلق بالأمر sem أدناه، والذي له مجموعة من الأسـئلة المنفصلة للمتغير *life* والمتغير *chldmort*

```
.sem (life<- adfert urban loggdp chldmort
 school reg1)
 (chldmort<- adfert urban loggdp school reg1),
 standardized
```

```
(16 observations with missing values excluded;
 specify option 'method(mlmv)' to use all observations)

Endogenous variables

Observed:  life chldmort

Exogenous variables

Observed:  adfert urban loggdp school reg1

Fitting target model:

Iteration 0:   log likelihood = -18660.532
Iteration 1:   log likelihood = -18660.532

Structural equation model                  Number of obs      =        978
Estimation method  = ml
Log likelihood     = -18660.532

------------------------------------------------------------------------------
             |                 OIM
Standardized |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  life <-    |
    chldmort |  -.4893582   .0285822   -17.12   0.000    -.5453783   -.4333381
      adfert |  -.0542246   .0176966    -3.06   0.002    -.0889092     -.01954
       urban |   .2355879   .0194596    12.11   0.000     .1974478    .2737279
      loggdp |   .0425882   .0257468     1.65   0.098    -.0078746    .0930509
      school |  -.0408058   .0212174    -1.92   0.054    -.0823912    .0007796
        reg1 |  -.2744216   .0216242   -12.69   0.000    -.3168041    -.232039
       _cons |   6.630625   .1913384    34.65   0.000     6.255609    7.005642
  -----------+----------------------------------------------------------------
  chldmort <-|
      adfert |   .1469444   .0187839     7.82   0.000     .1101287    .1837601
       urban |   .0136128   .0216136     0.63   0.529     -.028749    .0559746
      loggdp |  -.1766092   .0276417    -6.39   0.000    -.2307861   -.1224324
      school |   -.316167   .0205339   -15.40   0.000    -.3564127   -.2759212
        reg1 |   .4331934   .0186128    23.27   0.000      .396713    .4696738
       _cons |   2.695432   .1604064    16.80   0.000     2.381041    3.009822
-------------+----------------------------------------------------------------
Variance     |
      e.life |   .1570104   .0072756                      .1433788    .1719379
  e.chldmort |   .1900287   .0084368                      .1741919    .2073053
------------------------------------------------------------------------------
LR test of model vs. saturated: chi2(0)   =      0.00, Prob > chi2 =      .
```

التأثير غير المباشر والكلي يمكن حسابه بسهولة يدوياً، فالتأثير غير المباشر يساوي المُعاملات المتحصل عليها مع أي سلسلة للمسارات السببية التي تربط متغيرًا بآخر. التأثير الكلي يساوي مجموع كل التأثيرات المباشرة وغير المباشرة التي تربط متغيرين اثنين. عند قراءة المُعَامِلات المعيارية من

الشكل (15.8) فإنه يمكن القول بأن خصوبة المراهقين تؤثر على متوسط العمر المتوقع، وهذه التأثيرات كما يلي:

مباشر -0.05

غير مباشر $0.15\times(-0.49) = -0.07$

كلي $-0.05-0.07=-0.12$

أو بعبارة أخرى، هذا النموذج يتوقع أنه في حالة ثبات العوامل الأخرى، فإن زيادة الانحراف المعياري لمتغير خصوبة المراهقين بمقدار نقطة واحدة يؤدي إلى انخفاض الانحراف المعياري لمتوسط العمر المتوقع بمقدار 0.27 وذلك من خلال التأثيرات المباشرة وغير المباشرة. التأثير المباشر لمتغير *adfert* أقرب للصفر، ولكنه سببي مهم بسبب التأثير غير المباشر لمعدل وفيات الأطفال.

وبحسابات مشابهة لما قمنا به من قبل، فإن تأثير موقع أفريقيا (المشاكل الموجودة في أفريقيا لم يتم قياسها بأي متغير في النموذج) له تأثير أكثر من ضعف تأثير المتغير *reg1*:

مباشر -0.27

غير مباشر $0.43\times(-0.49) = -0.21$

كلي $-0.27-0.21=-0.48$

سوف نعود مرة أخرى لنماذج المعادلة الهيكلية في سياق التحليل العاملي factor analysis بالفصل (12).

الفصل التاسع

الانحدار اللوغاريتمي
Logistic Regression

طرق الانحدار التي سبق تناولها في الفصلين (7 و8) تتطلب – بشكل عام – متغيرات تابعة قابلة للقياس. ولكن برنامج ستاتا يوفر عددًا كبيرًا من التقنيات لصياغة النماذج التصنيفية، والترتيبية، والمتغيرات التابعة المقاربة. القائمة أدناه تعرض بعض الأفكار لعدة طرق متوافرة. وللحصول على تفاصيل حول أوامر معينة، قم بطباعة الأمر **help** *command*، وتعتبر دراسة Long and Freese (2006) من أفضل المصادر التي تشرح الطرق الرئيسة لبرنامج ستاتا للمتغيرات التابعة المحدودة، كما يمكنك الاطلاع على دراسة Hosmer and Lemeshow (2000).

asclogit بديل محدد بلوغاريتم مشروط (خيار ماك فادن McFadden).

asmprobit بديل محدد بانحدار الاحتمال المتعدد.

asroprobit بديل محدد بانحدار الاحتمال متعدد الرتب.

biprobit انحدار الاحتمال الثنائي.

binreg الانحدار الثنائي (نماذج خطية عامة).

blogit التقدير اللوغاريتمي مع بيانات مجمّعة (مقفلة).

bprobit تقدير احتمالي مع بيانات مجمّعة (مقفلة).

clogit الانحدار اللوغاريتمي للتأثيرات الثابتة المشروطة.

cloglog تقدير لوغاريتمي- لوغاريتمي مكمل.

constraint تعريف وتحديد واستبعاد القيود الخطية.

exlogistic الانحدار اللوغاريتمي الدقيق.

glm نماذج خطية عامة تتضمن خيارًا لإنشاء نموذج لوغاريتمي أو احتمالي أو روابط لوغاريتمية مكملة، مما يسمح لمتغير الاستجابة ليكون ثنائياً أو نسبياً لبيانات مجمعة.

glogit الانحدار اللوغاريتمي لبيانات مجمعة.

gprobit الانحدار الاحتمالي لبيانات مجمعة.

heckprob التقدير الاحتمالي مع الاختيار.

hetprob التقدير الاحتمالي لاختلاف التباين.

intreg انحدار الفترة، حيث إن y نقطة بيانات أوبيانات الفترة أو بيانات تم فحصها مسبقاً.

ivprobit الاحتمال مع متغيرات الانحدار الخارجية المستمرة.

logistic الانحدار اللوغاريتمي مع نسب الاحتمال.

logit الانحدار اللوغاريتمي وهو مشابه للأمر **logistic**، ولكن يعطي مُعامِلات بدلاً من نسب احتمال.

mlogit الانحدار اللوغاريتمي المتعدد للمتغيرات y متعددة التدريج.

nologit تقدير لوغاريتمي متداخل.

ologit الانحدار اللوغاريتمي لمتغيرات y الترتيبية.

oprobit الانحدار الاحتمالي لمتغيرات y الترتيبية.

probit الانحدار الاحتمالي لمتغير y الثنائية.

rologit النموذج اللوغاريتمي لتنظيم الرتب (كما يُعرف أيضاً بنموذج Plackett-Luce أو النموذج اللوغاريتمي الموسع أو تحليل المشترك الاختياري).

scobit التقدير الاحتمالي الملتوي.

slogit الانحدار اللوغاريتمي النمطي.

svy:logit الانحدار اللوغاريتمي مع بيانات دراسة استقصائية معقدة، كما يوجد كذلك العديد من أوامر الدراسات الاستقصائية **(svy)** الخاصة بنماذج المتغيرات التصنيفية. (لمزيد من التفاصيل قم بطباعة **help svy estimation**).

tobit انحدار توبت Tobit الذي يفترض أن *y* يتبع توزيع جاوس، ولكن يتم فحصه مسبقاً عند نقطة ثابتة معروفة. (للحصول على معلومات أكثر قم بطباعة الأمر **help cnreg**).

xtmelogit نماذج متعددة المستويات أو نماذج مختلطة لوغاريتمية ثنائية مع التأثيرات العشوائية أو الثابتة. للحصول على معلومات أكثر عن هذا الأمر، قم بطباعة **help xtmelogit** وسوف يتم شرح هذا الأمر لاحقاً في الفصل (13). وهناك العديد من الأوامر الخاصة بالبيانات الطولية (Panel Data). وللحصول على قائمة بهذه الأوامر قم بطباعة الأمر **help xt**.

بعد الاطلاع على أغلب أوامر النماذج، يمكن للأمر **predict** حساب القيم المتوقعة أو الاحتمالات. كما يمكن للأمر **predict** عرض الإحصائيات التشخيصية مثل تلك التي سبق شرحها للانحدار اللوغاريتمي في دراسة Hosmer and Lemeshow (2000)، استخدام خيار معين مع الأمر **predict** يعتمد على نوع النموذج المستخدم. هناك أمر يُستخدم بعد صياغة النموذج وهو الأمر **predictnl** يقوم بحساب التوقعات غير الخطية وفترات ثقتها. (لمزيد من المعلومات قم بطباعة الأمر **help predictnl**)، كما أن الأمر **margins** والأمر **marginsplot** من الأوامر المفيدة في هذا الصدد.

هناك العديد من الأمثلة عن هذه الأوامر سوف يتم تناولها لاحقاً في هذا الجزء. وعموماً فإن الطرق المتعددة للنماذج النوعية أو المتغيرات التابعة المحدودة، يمكن الوصول إليها عن طريق استخدام عدد من قوائم ستاتا. وهذه القوائم تتضمن:

Statistics > Binary outcomes
Statistics > Ordinal outcomes

Statistics > Categorical outcomes
Statistics > Generalized linear models
Statistics > Longitudinal/panel data
Statistics > Linear models and related
Statistics > Multilevel mixed-effects models

بعد الجزء التالي، سوف يتم التركيز في بقية هذا الفصل على مجموعة مهمة من الطرق تسمى الانحدار اللوغاريتمي أو اللوجيستي. سوف نقوم بإلقاء نظرة سريعة على الطرق اللوغاريتمية الأساسية للمتغيرات التابعة ذات التصنيفات المتعددة والترتيبية والثنائية. الفصل (13) يقوم بشرح الشكل اللوغاريتمي لنماذج التأثيرات الثابتة.

أمثلة عن الأوامر : Example Commands

```
.logistic y x1 x2 x3
```

يقوم بحساب الانحدار اللوغاريتمي للمتغير الثنائي $y\{0,1\}$ على المتغيرات التنبؤية $x3, x2, x1$ وتأثيرات المتغير التنبؤي يتم عرضها كنسب احتمالية، والأمر الأقرب للأمر أعلاه هو **logit** الذي يقوم بحساب نفس الانحدار تقريباً، ولكن يعرض التأثيرات كمُعَامِلات انحدار محتملة مسجلة. والنماذج المحددة بالأمر **logistic** والأمر **logit** هي نفسها، ولذلك فإن التوقعات والاختبارات التشخيصية سوف تكون متطابقة.

```
.estat gof
```

يقوم هذا الأمر بحساب اختبار مربع كاي لبيرسون لحُسن المطابقة للنموذج اللوغاريتمي وهو عبارة عن: مقارنة التكرارات المتوقعة مع المحسوبة للمتغير $y = 1$ باستخدام خلايا تم تعريفها بواسطة نمط المتغيرات المستقلة (المتغير x)، عند وجود عدد كبير من أنماط المتغير x فيمكننا وضع هذه الأنماط في مجموعات حسب الاحتمالات المقدّرة، الأمر **estat gof, group(10)** يقوم بحساب الاختبار مع 10 مجموعات متساوية تقريباً.

```
.estat classification
```

يقوم بعرض الإحصائيات المصنفة وجدول التصنيفات، الأمر **estat classification** والأمر **lroc** والأمر **lsens** (المعروضة أدناه) تعتبر مفيدة عندما يكون هدف التحليل هو التصنيف. هذه الأوامر جميعها تشير إلى النموذج اللوغاريتمي السابق.

```
.lroc
```

يقوم بإنشاء رسم بياني يوضح منحنى خاصية عمل المتلقي (ROC) ويقوم بحساب المنطقة التي تقع تحت المنحنى.

```
.lsens
```

يقوم بإنشاء رسم بياني للحساسية والخصوصية مقابل نقطة القطع الاحتمالي.

```
.predict phat
```

يقوم بإنشاء متغير جديد (تم تسميته عشوائياً باسم *phat*) يساوي الاحتمالات المتوقعة التي فيها المتغير $y = 1$ وذلك بناءً على آخر نموذج لوغاريتمي.

```
.predict dX2, dx2
```

يقوم بإنشاء متغير جديد (تم تسميته عشوائياً باسم *dX2*)، الإحصائية التشخيصية تقوم بقياس التغير في اختبار مربع كاي لبيرسون من آخر تحليل لوغاريتمي.

```
.mlogit y x1 x2 x3, base(3) rrr nolog
```

يقوم بحساب الانحدار اللوغاريتمي المتعدد لمخرجات متعددة للمتغير y على ثلاثة متغيرات x، الخيار **base(3)** يقوم بتحديد أن $y = 3$ كتصنيف أساسي للمقارنة، والخيار **rrr** يقوم بعرض نسب الخطر المتعلقة بمُعَامِلات الانحدار، والخيار **nolog** يمنع عرض سجل الاحتمال لكل تكرار.

```
.svy:mlogit y x1 x2 x3, base(3) rrr nolog
```

يقوم بحساب الانحدار اللوغاريتمي المتعدد الموزون للدراسات الاستقصائية، وهذا يتطلب أن تكون البيانات قد تم تعريفها مسبقاً على أنها بيانات دراسة استقصائية باستخدام الأمر **svyset** (انظر الفصل 4)، الأشكال المختلفة لأوامر الدراسات الاستقصائية، وأوامر الانحدار اللوغاريتمي

logit, ologit وأوامر صياغة النماذج الأخرى لها نفس التركيبة تقريباً وتشبه نظيراتها الأخرى.

```
.predict P2, outcome(2)
```

يقوم بإنشاء متغير جديد (تم تسميته عشوائياً باسم *P2*) يمثل الاحتمال المتوقع عندما $y = 2$ بناءً على آخر تحليل تم إجراؤه بالأمر **mlogit**.

```
.glm successx1 x2 x3, family(binomial trials)
 eform
```

يقوم هذا الأمر بحساب الانحدار اللوغاريتمي من خلال إنشاء نموذج خطي عام باستخدام جداول بدلاً من المشاهدات الموجودة بالبيانات. المتغير *success* يعطي عدد المرات التي ظهرت فيها المخرجات ذات العلاقة. أما المتغير *trials* فيعطي عدد المرات التي كان يمكن أن تحدث لكل مجموعة من المتغيرات التنبؤية *x3, x2, x1*. ولذا فإن المتغيرين *success/trials* سوف يساويان عدد المرات التي حدثت فيها المخرجات مثل احتمال تعافي المريض، الخيار **eform** يعرض النتائج في شكل نسب مرجحة ("شكل أُسّي") بدلاً من مُعامِلات لوغاريتمية.

بيانات مكوك الفضاء : Space Shuttle Data

المثال الأول في هذا الفصل، يتضمن بيانات موجودة بالملف *shuttle.dta*، وهو يحتوي على بيانات تاريخية تغطي أول 25 رحلة لمكوك الفضاء الأمريكي. هذه البيانات تتضمن دليل أنه إذا تم تحليل البيانات بطريقة صحيحة، فإن النتائج يجب أن تُظهِر بأن موظفي وكالة ناسا يُفترض أنهم لم يُطلِقوا مكوك الفضاء تشالنجر في رحلته المميتة في سنة 1985 (الرحلة 25 لمكوك الفضاء والتي كانت تحمل رقم STS 51-L). البيانات تم الحصول عليها من تقرير رئيس لجنة التحقيق في حادثة مكوك الفضاء تشالنجر في سنة 1986، ومن كتاب Tufte (1997)، هذا الكتاب يتضمن نقاشاً رائعاً حول البيانات وقضايا التحليل. تعليقات Tufte بشأن رحلات فضائية محددة تتضمن متغيرًا نصّياً في هذه البيانات.

```
.use C:\data\shuttle.dta, clear
.describe
```

```
Contains data from C:\data\shuttle.dta
  obs:            25                        First 25 space shuttle flights
 vars:             8                        2 Jul 2012 06:11
 size:         1,575

-----------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-----------------------------------------------------------------------------
flight          byte   %8.0g       flbl       Flight
month           byte   %8.0g                  Month of launch
day             byte   %8.0g                  Day of launch
year            int    %8.0g                  Year of launch
distress        byte   %8.0g       dlbl       Thermal distress incidents
temp            byte   %8.0g                  Joint temperature, degrees F
damage          byte   %9.0g                  Damage severity index, Tufte 1997
comments        str55  %55s                   Comments, Tufte 1997
-----------------------------------------------------------------------------
Sorted by:  flight
```

.list *flight-temp*, sepby(*year*)

	flight	month	day	year	distress	temp
1.	STS-1	4	12	1981	none	66
2.	STS-2	11	12	1981	1 or 2	70
3.	STS-3	3	22	1982	none	69
4.	STS-4	6	27	1982	.	80
5.	STS-5	11	11	1982	none	68
6.	STS-6	4	4	1983	1 or 2	67
7.	STS-7	6	18	1983	none	72
8.	STS-8	8	30	1983	none	73
9.	STS-9	11	28	1983	none	70
10.	STS_41-B	2	3	1984	1 or 2	57
11.	STS_41-C	4	6	1984	3 plus	63
12.	STS_41-D	8	30	1984	3 plus	70
13.	STS_41-G	10	5	1984	none	78
14.	STS_51-A	11	8	1984	none	67
15.	STS_51-C	1	24	1985	3 plus	53
16.	STS_51-D	4	12	1985	3 plus	67
17.	STS_51-B	4	29	1985	3 plus	75
18.	STS_51-G	6	17	1985	3 plus	70
19.	STS_51-F	7	29	1985	1 or 2	81
20.	STS_51-I	8	27	1985	1 or 2	76
21.	STS_51-J	10	3	1985	none	79
22.	STS_61-A	10	30	1985	3 plus	75
23.	STS_61-B	11	26	1985	1 or 2	76
24.	STS_61-C	1	12	1986	3 plus	58
25.	STS_51-L	1	28	1986	.	31

هذا الفصل، يقوم باختبار ثلاثة متغيرات بملف البيانات *shuttle.dta* هي:

distress يمثل عدد "حوادث التلف الحراري" التي انفجر فيها الغاز الساخن أو تلف حراري في أغلفة الوصلات لصواريخ الدفع، هذا التلف في أغلفة الوصلات في صواريخ الدفع ساهمت بدرجة كبيرة بحادثة تشالنجر، وقد عانت العديد من الرحلات الفضائية السابقة من تلف أقل خطورة، لذلك كان من المعروف أن أغلفة الوصلات هي مصدر خطر محتمل.

temp درجة حرارة الوصلات المحسوبة عند الإطلاق مقاسة بالفهرنهايت. درجة الحرارة تعتمد بشكل كبير على درجة حرارة الجو، حيث إن وصلات صواريخ الإطلاق المطاطية الدائرية تصبح أقل مرونة في البرودة.

date متغير التاريخ وهي مقاسة بالأيام التي انقضت بعد 1 يناير 1960، المتغير *date* عبارة عن شهر ويوم وسنة الإطلاق باستخدام الدالة **mdy** (شهر – يوم – سنة للزمن المنقضي؛ لمزيد من التفاصيل قم بطباعة **help dates**).

```
.generate date = mdy(month, day, year)
.format %td date
.label variable date "Date (days since 1/1/60)"
```

التواريخ المنقضية مهمة، لأن العديد من التغيرات خلال فترة برنامج الرحلات الفضائية ربما أصبحت أكثر خطورة بمضي الزمن. حيث إن حوائط صاروخ الإطلاق أصبحت أقل سُمكاً لتوفير مساحة وزيادة الحمولة. وإغلفة الوصلات تم اختبارها تحت ضغط أعلى. بالإضافة إلى ذلك، فإن أجهزة المكوك الفضائي القابلة لإعادة الاستخدام تم صيانتها، لذلك فإننا قد نتساءل: هل احتمال أن التلف المصاحب لصواريخ الدفع (واحد أو أكثر حادثة تلف حراري) زادت مع تواريخ الإطلاق؟

المتغير *distress* تم توصيفه كمتغير رقمي.

```
.tabulate distress
```

Thermal distress incidents	Freq.	Percent	Cum.
none	9	39.13	39.13
1 or 2	6	26.09	65.22
3 plus	8	34.78	100.00
Total	23	100.00	

في العادة، فإن الأمر **tabulate** يقوم بعرض توصيف القــيم، والخيــار **nolabel** يوضح أن الرموز الرقمية التي تصف القيم هي "none" = 0 ، "1 or 2" = 1 ، "3 plus" = 2

```
.tabulate distress, nolabel
```

Thermal distress incidents	Freq.	Percent	Cum.
0	9	39.13	39.13
1	6	26.09	65.22
2	8	34.78	100.00
Total	23	100.00	

يمكننا استخدام هذه الرموز لإنشاء متغير وهمي باسم *any*، هذا المتغير الذي يساوي 0، يمثل عدم حدوث أي مشاكل حرارية، و1 عند حدوث مشكلة حرارية واحدة أو أكثر كما يلي:

```
.generate any = distress
.replace any = 1 if distress == 2
.label variable any "Any thermal distress"
```

لفحص ما إذا كان الأمر **generate** والأمر **replace** تــم إنجازهمــا كمــا يجب، والتأكد بأن القيم المفقودة تم التعامل معها بشكل صحيح، قم بطباعــة الأمر:

```
.tabulate distress any, miss
```

Thermal distress incidents	Any thermal distress 0	1	.	Total
none	9	0	0	9
1 or 2	0	6	0	6
3 plus	0	8	0	8
.	0	0	2	2
Total	9	14	2	25

نماذج الانحدار اللوغاريتمي تحدد كيف أن متغيرًا ثنائياً {0,1} مثل متغير *any* يعتمد على واحد أو أكثر من متغيرات *x*. تركيبة الأمر **logit** تشبه تركيبة الأمر **regress**، وأغلب أوامر النماذج الأخرى مع متغير تابع يكون في الأول.

```
.logit any date
Iteration 0:   log likelihood = -15.394543
Iteration 1:   log likelihood = -12.997472
Iteration 2:   log likelihood = -12.991097
Iteration 3:   log likelihood = -12.991096

Logistic regression                         Number of obs   =         23
                                            LR chi2(1)      =       4.81
                                            Prob > chi2     =     0.0283
Log likelihood = -12.991096                 Pseudo R2       =     0.1561
```

any	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
date	.0020907	.0010703	1.95	0.051	-6.94e-06	.0041884
_cons	-18.13116	9.517253	-1.91	0.057	-36.78463	.5223142

الأمر **logit** يكرر طريقة التقدير للوصول إلى الحد الأعلى من سجل دالة الاحتمال المعروضة في الجدول اللوغاريتمي أعلاه. عند التكرار 0، فإن سجل الاحتمال يشرح تناسب النموذج متضمناً ثابتاً فقط. آخر سجل للاحتمال يشرح تناسب النموذج النهائي.

$$L = -18.13116 + 0.0020907 date \quad [9.1]$$

حيث إن L يمثل اللوغاريتم المتوقع أو سجل الترجيح لأي حادث تلف حراري:

$$L = \ln[P(any = 1) / P(any = 0)] \quad [9.2]$$

واختبار χ^2 عموماً في الجانب الأيمن العلوي يقوم بتقييم فرضية العدم التي تفترض أن كل المُعَامِلات في النموذج تساوي صفرًا باستثناء الثابت.

$$\chi^2 = -2(\ln\mathscr{L}_i - \ln\mathscr{L}_f) \quad [9.3]$$

حيث إن $\mathscr{L}$ تمثل الاحتمال المسجل الأولى أو التكراري 0 (نموذج مع ثابت فقط) و$\ln\mathscr{L}_f$ تمثل الاحتمال المسجل للتكرار النهائي، حيث يمكن كتابة المعادلة كما يلي:

$$\chi^2 = -2[-15.394543-(-12.991096)]$$
$$=4.81$$

احتمال الحصول على قيمة أكبر لــ χ^2 مع درجة حرية تساوي 1 (الفرق في التعقيد بين النماذج الأولية والنهائية) احتمال منخفض بما فيه الكفاية (0.0283) لرفض فرضية العدم في هذا المثال. وبالتالي، فإن المتغير *data* ليس له تأثير ذو معنوية إحصائية.

بالرغم من انخفاض دقتها وسهولتها، فإن الاختبارات التي يوفرها z (الاعتدال المعياري) إحصائيات تم عرضها مع نتائج **logit**، مع وجود متغير تنبؤي واحد وهو إحصائية z للمتغير التنبؤي واحصائية χ^2 تختبر الفرضيات المكافئة، فإن هذا الوضع يشبه تماماً إحصائيات اختبارات F وt في انحدار OLS البسيط، وعلى خلاف نظائر OLS فإن تقريب z اللوغاريتمي، واختبارات χ^2 أحياناً لا تتفق (وكذلك نحن)، اختبار χ^2 يكون صالحاً بشكل عام.

مشابهاً لبعض طرق الاحتمال ببرنامج ستاتا، فإن الأمر **logit** يعرض محدد R^2 وهمي كما يلي:

$$R^2_{\text{الوهمي}} = 1 - \ln\mathscr{L}_f / \ln\mathscr{L}_i \quad [9.4]$$

وفي هذا المثال، فإن المعادلة أعلاه ستكون:

$$R^2_{\text{الوهمي}} = 1 - (-12.991096) / (-15.394543)$$
$$= 0.1561$$

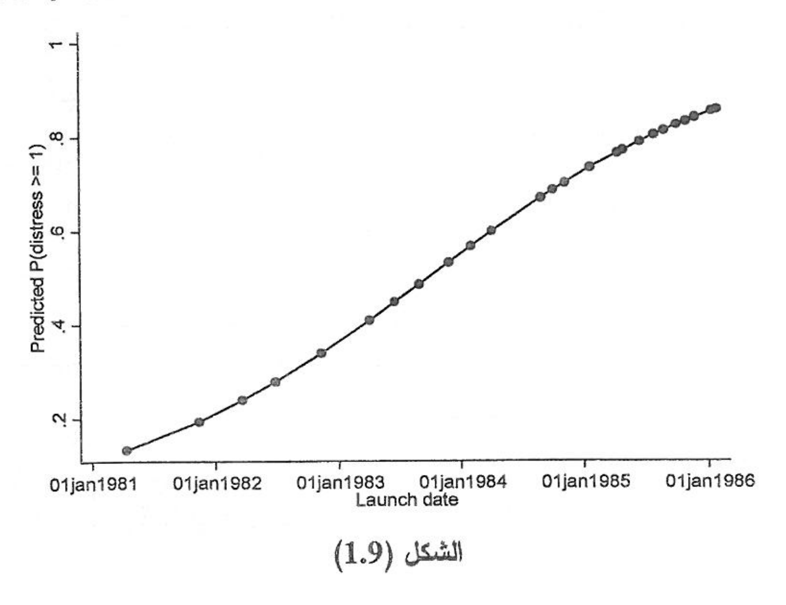
بالرغم من أن إحصائيات R^2 الوهمية تعتبر طريقة سريعة لوصف ومقارنة مدى تناسب النماذج المختلفة لنفس المتغير التابع، إلا أنها تفتقر إلى التفسير الواضح للتباين بالنسبة R^2 الحقيقية في تحليل انحدار OLS.

بعد التحليل باستخدام الأمر **logit**، فإن استخدام الأمـر **predict** (بـدون إضافة أي خيارات) يُمكّننا من الحصول على الاحتمالات المتوقعة.

$$Phat = 1 / (1 + e^{-L}) \qquad [9.5]$$

عند إنشاء رسم بياني للاحتمالات المتوقعة مـع المتغيـر *date*، فـإن المنحنى اللوغاريتمي سوف يكون على شكل حرف S، لأننا قمنـا بتحديـد الخيارات **format %td** ***date*** سابقاً بعد أن قمنا بتعريف المتغير *date*، القيم تم توصيفها بشكل مناسب في المحور الأفقي أو محور الزمن في الشكل (1.9).

```
.predict Phat
.label variable Phat "Predicted P(distress >= 1)"
.graph twoway connect Phat date, xtitle("Launch
 date") sort
```



الشكل (1.9)

مُعَامِل الأمر **logit** في هذا المثال (0.0020907) يصف تأثير المتغيـر *date* على الاحتمالات اللوغاريتمية لوقوع حوادث التلف الحراري، كل يـوم إضافي يزيد من الاحتمالات اللوغاريتمية المتوقعة للتلف الحـراري بقيمـة 0.0020907 وبعبارة أخرى يمكننا القول إن كل يـوم إضـافي يـضاعف الاحتمالات اللوغاريتمية المتوقعة للتلف الحراري بقيمـة $(e^{0.0020907})^{100}=1.23$، ($e\approx$ 2.71828 الرقم الأساسي للوغاريتم الطبيعي)، ويمكن لبرنامج ستاتا حفظ المُعَامِلات بعد كل عملية تحليل، وذلك بإضافة _b[*varname*] كما يلي:

```
.display exp(_b[date])
 1.0020929
.display exp(_b[date])^100
 1.2325358
```

يمكننا ببساطة إضافة الخيار **or** (النسبة الاحتمالية) مع سطر الأمر **logit**، الطريقة الثالثة للحصول على النسب الاحتمالية تتم عن طريق الأمر **logistic**، وسيتم شرحها في الجزء التالي. الأمر **logistic** يتناسب بالضبط مع نفس النماذج التي يتناسب معها الأمر **logit**، ولكن الوضع الافتراضي لجدول مخرجاته يعرض النسب الاحتمالية بدلاً من عرض المُعَامِلات.

استخدام الانحدار اللوغاريتمي : Using Logistic Regression

الأمر أدناه هو نفس تحليل الانحدار الذي قمنا به سابقاً، ولكن باستخدام الأمر **logistic** بدلاً من الأمر **logit**.

```
.logistic any date
```

```
Logistic regression                               Number of obs     =         23
                                                  LR chi2(1)        =       4.81
                                                  Prob > chi2       =     0.0283
Log likelihood = -12.991096                       Pseudo R2         =     0.1561
```

any	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
date	1.002093	.0010725	1.95	0.051	.9999931	1.004197
_cons	1.34e-08	1.27e-07	-1.91	0.057	1.06e-16	1.685925

يجب ملاحظة تشابه الاحتمالات اللوغاريتمية وإحصائيات χ^2، وبدلاً من عرض المُعَامَلات فإن الأمر **logistic** يعرض النسب الاحتماليــة (e^b)، القــيم المعروضة بالعمود Odds Ratio لجدول مخرجات الأمر **logistic** أعلاه هــي عبارة عن أرقام، وهذه الأرقام هي احتمالات وكل زيادة في الاحتمال بمقدار وحدة واحدة يؤدي إلى زيادة في المتغير x (إذا كانت قيم متغيرات x الأخرى ثابتة).

بعد تحديد النموذج المناسب يمكننا الحصول علــى جــدول تــصنيفي والإحصائيات ذات العلاقة بطباعة الأمر التالي:

```
.estat class
```

```
Logistic model for any

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |        12             4  |         16
     -     |         2             5  |          7
-----------+--------------------------+-----------
   Total   |        14             9  |         23

Classified + if predicted Pr(D) >= .5
True D defined as any != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   85.71%
Specificity                     Pr( -|~D)   55.56%
Positive predictive value       Pr( D| +)   75.00%
Negative predictive value       Pr(~D| -)   71.43%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   44.44%
False - rate for true D         Pr( -| D)   14.29%
False + rate for classified +   Pr(~D| +)   25.00%
False - rate for classified -   Pr( D| -)   28.57%
--------------------------------------------------
Correctly classified                        73.91%
--------------------------------------------------
```

الوضع الافتراضي للأمر **estat class** أن يقوم باستخدام احتمال 0.5 كنقطة قطع (ويمكننا تغييره بإضافة الخيار **cutoff()**)، الرموز التي تظهر في جدول التصنيف أعلاه لها المعاني التالية:

- D وقوع الحدث موضع الدراسة (وهذا يعني $y = 1$) لتلك المشاهدة، وفي هذا المثال D تشير إلى وقوع تلف حراري.
- ~D عدم وقوع الحدث موضع الدراسة (وهذا يعني $y = 0$) لتلك المشاهدة، وفي هذا المثال D ~ ترتبط بالرحلات التي لم يحدث فيها تلف حراري.
- + الاحتمال المتوقع للنموذج أكبر من أو يساوي نقطة القطع، وحيث إننا استخدمنا نقطة القطع الافتراضية، فإن علامة + هنا تشير إلى أن النموذج يتوقع 0.5 أو أكبر لاحتمال التلف الحراري.
- − الاحتمال المتوقع أقل من نقطة القطع، وهنا علامة − تعني أن الاحتمال المتوقع للتلف الحراري أقل من 0.5.

ولذلك فإن 12 رحلة وتصنيفاتها كانت دقيقة في سياق أن النموذج قام بتوقّع 0.5 من احتمال التلف الحراري وهذا وقع بالفعل. وبالنسبة لبقية 5 رحلات، فإن النموذج يتوقع أقل من 0.5 من احتمال التلف الحراري وهذا لم يحدث فعلاً. وبصفة عامة، فإن معدل التصنيف الصحيح يكون 12+5 = 17 رحلة من أصل 23 رحلة أو 73.91%، كما أن الجدول يعرض الاحتمالات المشروطة مثل الحساسية أو نسبة المشاهدات التي تكون فيها $P \geq 0.5$ مع ملاحظة حدوث التلف الحراري (12 من أصل 14 أو 85.71%).

بعد استخدام الأمر **logistic** أو الأمر **logit** يمكننا استخدام الأمر **predict** لحساب توقعات متعددة وإحصائيات تشخيصية أخرى، يمكن الحصول على شرح عن النماذج اللوغاريتمية والإحصائيات التشخيصية في دراسة Hosmer and Lemeshow (2000).

predict ***newvar***	الاحتمال المتوقع الذي يكون فيه $y = 1$
predict ***newvar*, xb**	الاحتمال الخطي (احتمالات لوغاريتمية متوقعة يكون فيها $y = 1$)
predict ***newvar*, stdp**	الخطأ المعياري للتنبؤ الخطي.
predict ***newvar*, dbeta**	ΔB تأثير الإحصائيات وهو يماثل مسافة كوك Cook's D
predict ***newvar*, deviance**	الانحراف المتبقي للمشاهدة jth لنمط المتغير x وهو d_j
predict ***newvar*, dx2**	التغير في χ^2 لبيرسون والذي يمكن كتابته $\Delta\ \chi^2$ أو $\Delta\ \chi^2_P$
predict ***newvar*, ddeviance**	التغير في انحراف χ^2 ويمكن كتابته ΔD أو $\Delta\ \chi^2_D$
predict ***newvar*, hat**	تأثير المشاهدة jth على نمط المتغير x ويكون هو h_j
predict ***newvar*, number**	تخصيص أرقام لأنماط المتغير x حيث تكون $j = 1, 2, 3 \ldots J$

predict *newvar*, resid	باقي بيرسون للمشاهدة *j*th لـنمط المتغير *x* وهذا الباقي هو r_j
predict *newvar*, rstandard	باقي بيرسون المعياري
predict *newvar*, score	أول مشتقة من الاحتمال اللوغاريتمي مع الأخذ بالاعتبار **Xb**

الإحصائيات التي يتم الحصول عليها عن طريق استخدام الخيارات **dbeta, dx2, ddeviance, hat** لا تقوم بقياس تأثير كل مشاهدة على حدة كما تفعل نظائرها في تحليل الانحدار العادي، وإنما تقوم بقياس تأثير أنماط المتغيرات المستقلة، وبالتالي فهي تستبعد المشاهدات التي تحتوي على مجموعة خاصة من قيم *x*. لمزيد من المعلومات انظر دراسة Hosmer and Lemeshow (2000).

هل صواريخ الدفع مع درجة الحرارة معاً يؤثران على احتمال حدوث التلف الحراري؟ يمكننا التحقق من ذلك عن طريق تضمين متغير درجة الحرارة *temp* كمتغير تنبؤي ثان.

```
.logistic any date temp
```

```
Logistic regression                               Number of obs   =         23
                                                  LR chi2(2)      =       8.09
                                                  Prob > chi2     =     0.0175
Log likelihood = -11.350748                       Pseudo R2       =     0.2627
```

any	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
date	1.00297	.0013675	2.17	0.030	1.000293	1.005653
temp	.8408309	.0987887	-1.48	0.140	.6678848	1.058561
_cons	1.19e-06	.0000121	-1.34	0.182	2.40e-15	587.9723

تضمين درجة الحرارة كمتغير تنبؤي طوّر بشكل جزئي معدل التصنيف الصحيح ليكون 78.26%.

```
.estat class
```

```
Logistic model for any

              ———— True ————
Classified |        D          ~D  |     Total
-----------+-----------------------+----------
     +     |       12           3  |        15
     -     |        2           6  |         8
-----------+-----------------------+----------
   Total   |       14           9  |        23

Classified + if predicted Pr(D) >= .5
True D defined as any != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   85.71%
Specificity                     Pr( -|~D)   66.67%
Positive predictive value       Pr( D| +)   80.00%
Negative predictive value       Pr(~D| -)   75.00%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   33.33%
False - rate for true D         Pr( -| D)   14.29%
False + rate for classified +   Pr(~D| +)   20.00%
False - rate for classified -   Pr( D| -)   25.00%
--------------------------------------------------
Correctly classified                        78.26%
--------------------------------------------------
```

حسب النموذج أعلاه، فإن زيادة درجة الحرارة درجة مئوية واحدة يُضاعف احتمالات تلف صواريخ الدفع بمقدار 0.84، بعبارة أخرى، كل نقص في درجة الحرارة بمقدار درجة مئوية واحدة يقلل احتمالات التلف بمقدار 16%؛ بالرغم من أن هذا التأثير يبدو قوياً بما فيه الكفاية ليُثير الاهتمام، فإن اختبار z يوضح بأن هذا الاحتمال ليس ذا معنوية إحصائية (z=-1.476, p=0.140)، والاختبار الآخر الحاسم يتم باستخدام نسبة الاحتمال لاختبار χ^2، الأمر **lrtest** يقوم بمقارنة نماذج متداخلة يتم تقديرها باستخدام الأرجحية العظمى، حيث يتم أولاً تقدير النموذج بالكامل بما في ذلك كل المتغيرات ذات العلاقة كما تم فعله سابقاً مع الأمر **logistic *any date temp*** ثم بعد ذلك نقوم بطباعة الأمر **estimates store** يليه اسم (بافتراض أن الاسم هو *full*) لتحديد النموذج الأول:

```
.estimates store full
```

الآن نقوم بتقدير النموذج المصغّر والذي يتضمن فقط مجموعة من متغيرات x التي توجد في النموذج الكامل أعلاه (يُطلق على مثل هذه النماذج

المصغّرة اسم النماذج المتداخلة)، وأخيراً استخدام أمر مثل lrtest *full* يقوم بإجراء اختبار للنموذج المتداخل مع النموذج السابق والذي قمنا بتسميته *full*. فعلى سبيل المثال، (قمنا باستخدام **quietly** قبل الأمر **logistic** أدناه، لأننا شاهدنا المخرجات من قبل).

```
.quietly logistic any date
.lrtest full
```

```
Likelihood-ratio test                                 LR chi2(1)  =      3.28
(Assumption: . nested in full)                        Prob > chi2 =    0.0701
```

الأمر **lrtest** يقوم باختبار آخر (يُفترض أنه متداخل) نموذج مع النموذج الذي سبق حفظه بواسطة الأمر **estimates store** وهو يستخدم إحصائية اختبار عامة لنماذج الأرجحية العظمى المتداخلة.

$$\chi^2 = -2(\ln \mathscr{L}_1 - \ln \mathscr{L}_0) \qquad [9.6]$$

حيث إن $\mathscr{L}_0$ ln الاحتمال اللوغاريتمي المتوقع للنموذج الأول (مع كل متغيرات x)، $\mathscr{L}_1$ ln الاحتمال اللوغاريتمي المتوقع للنموذج الثاني (مع مجموعة محددة من متغيرات x فقط)، قارن إحصائيات الاختبار الناتجة مع توزيع χ^2 مع درجات حرية تساوي الفرق في التعقيد (عدد متغيرات x المستبعدة) بين النماذج 1 و0، لمزيد من المعلومات حول هذا الأمر قم بطباعة **help lrtest** ، وهذا الأمر يعمل مع أي حسابات لتقدير الأرجحية العظمى ببرنامج ستاتا (هذه الأوامر مثل **logit, mlogit, stcox** والعديد من الأوامر الأخرى). وبصفة عامة، فإن إحصائية χ^2 يتم حسابها باستخدام مخرجات الأمر **logit** أو الأمر **logistic** (المعادلة رقم [9.3]) وخصوصاً المعادلة رقم [9.6].

الأمر **lrtest** في المثال السابق يقوم بإجراء العملية الحسابية التالية:

$$\chi^2 = -2[-12.991096 - (-11.350748)]$$
$$=3.28$$

مع درجة حرية 1 سوف يكون لدينا $p = 0.0701$، وتأثير المتغير *temp* ذي معنوية إحصائية عند $\alpha = 0.10$، وبالأخذ في الاعتبار أننا نقوم باستخدام

عينة صغيرة الحجم والتأثيرات السلبية المحتملة للخطأ من النوع الثاني بخصوص السلامة بالمركبة الفضائية فإن 0.10 = α يبدو أنها نقطة تحول أكثر دقة من المعتاد عندما تكون 0.05 = α.

الرسم البياني للتأثيرات المشروطة أو الهامشية :

Marginal or Conditional Effects Plots

إنشاء رسم بياني للتأثيرات المشروطة أو الهامشية المعدلة يساعد في فهم وتوصيل معنى النموذج اللوغاريتمي وتطبيقاته على الاحتمالات. فمثلاً يمكننا حساب الاحتمال المتوقع لحوادث التلف الحراري كدالة للمتغير *temp* مع الحفاظ على المتغير *date* عند قيم منخفضة (في البداية) أو مرتفعة (في النهاية) نسبياً.

```
.summarize date temp
```

Variable	Obs	Mean	Std. Dev.	Min	Max
date	25	8905.88	517.6033	7772	9524
temp	25	68.44	10.52806	31	81

التواريخ الماضية في هذا المدى من البيانات من أول رحلة للمكوك الفضاء في 12 أبريل 1981 (*date* = 7772) إلى تاريخ كارثة تشالنجر في 21 يناير 1986 (*date* = 9524)، درجات الحرارة تتراوح ما بين 31 إلى 81 درجة فهرنهايتية، الأمر **margins** يمكنه إجراء الحسابات، بينما الأمر **marginsplot** يمكنه إنشاء الأشكال البيانية للتوقعات المحتملة للنموذج الذي تم الحصول عليه من الأمر **logistic** عند زيادة درجة الحرارة 10 درجات عند تواريخ البداية والنهاية.

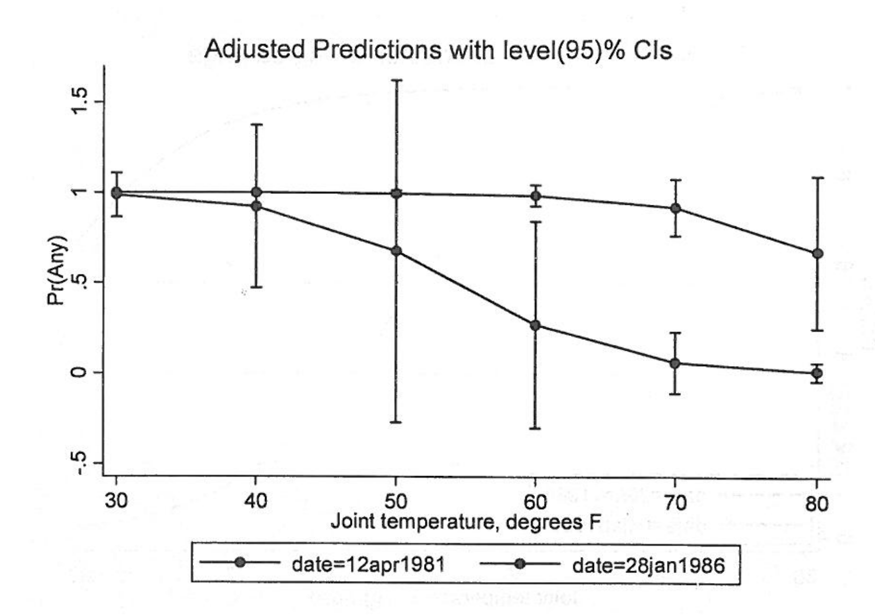
```
.quietly logistic any date temp
.margins, at(temp = (30(10)80) date = (7772
 9524)) vsquish
```

```
Adjusted predictions                              Number of obs     =         23
Model VCE    : OIM

Expression   : Pr(any), predict()
1._at        : date            =        7772
               temp            =          30
2._at        : date            =        7772
               temp            =          40
3._at        : date            =        7772
               temp            =          50
4._at        : date            =        7772
               temp            =          60
5._at        : date            =        7772
               temp            =          70
6._at        : date            =        7772
               temp            =          80
7._at        : date            =        9524
               temp            =          30
8._at        : date            =        9524
               temp            =          40
9._at        : date            =        9524
               temp            =          50
10._at       : date            =        9524
               temp            =          60
11._at       : date            =        9524
               temp            =          70
12._at       : date            =        9524
               temp            =          80
```

	Margin	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
_at						
1	.985239	.0624661	15.77	0.000	.8628077	1.10767
2	.9218137	.2310152	3.99	0.000	.4690321	1.374595
3	.6755951	.4831333	1.40	0.162	-.2713288	1.622519
4	.2689325	.291999	0.92	0.357	-.3033749	.84124
5	.0610143	.0871295	0.70	0.484	-.1097563	.2317849
6	.0113476	.0255353	0.44	0.657	-.0387006	.0613958
7	.9999169	.0004545	2200.25	0.000	.9990262	1.000808
8	.99953	.0020277	492.94	0.000	.9955558	1.003504
9	.9973449	.0084046	118.67	0.000	.9808722	1.013818
10	.9851528	.0302581	32.56	0.000	.9258479	1.044458
11	.9213867	.0808455	11.40	0.000	.7629324	1.079841
12	.6742985	.2166156	3.11	0.002	.2497398	1.098857

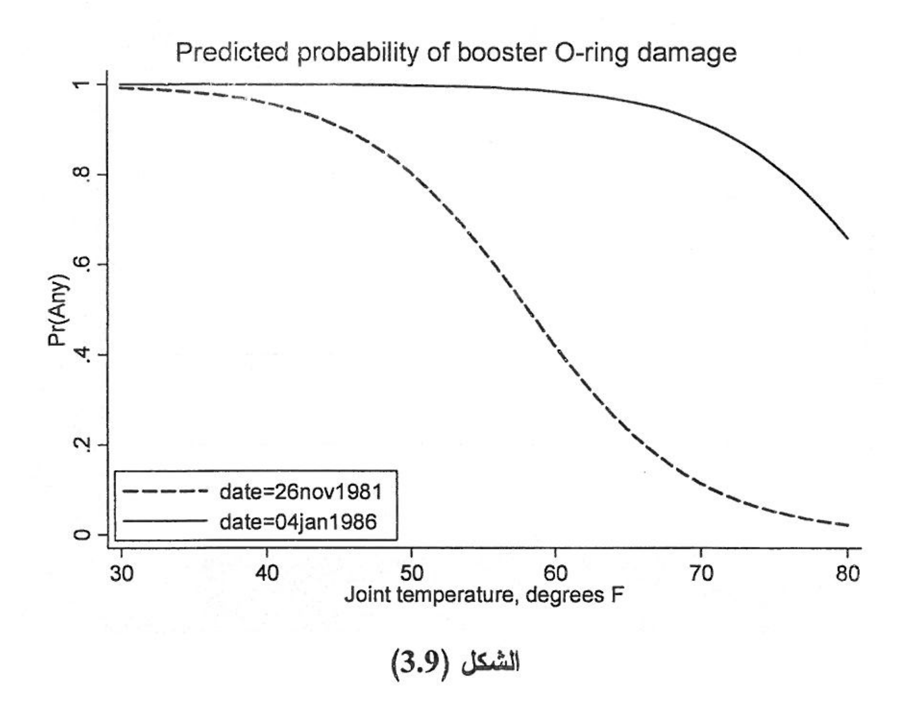
.marginsplot



الشكل (2.9)

الوضع الافتراضي للأمر **marginsplot** بالشكل (2.9) يوضح المعلومات الأساسية، ولكنه ليس بالتنسيق المطلوب. وللحصول على شكل أكثر قبـولاً للنشر، يمكننا منع ظهور فترات الثقة عن طريق الخيار **noci** ونقـل مربـع شرح الرسم واستخدام الخيار **plot#opts()** لإظهار المنحنيين وإضافة عنوان، أولاً سوف نقوم باستخدام الأمر **margins** مع زيادة درجة الحـرارة درجـة واحدة، وهذه الزيادة سوف تجعل المنحنيات الناتجة أكثر تجانساً.

```
.quietly margins, at(temp = (30(1)80) date =
 (8000 9500))
.marginsplot, noci legend(position(7) ring(0)
 rows(2))
 plot1opts(msymbol(i) lpattern(dash)
 lwidth(medthick))
 plot2opts(msymbol(i) lpattern(solid)
 lwidth(medthich))
 title("Predicted probability of booster O-ring
 damage")
```



الشكل (3.9)

بناءً على نموذجنا اللوغاريتمي، فإنه بالقرب من وقت البداية لرحلة مكوك الفضاء (المنحنى المتقطع في الشكل أعلاه) احتمال التلف الحراري يتجه من القرب من الصفر عند درجة حرارة 80 °F إلى أقل من 40 °F تقريباً، وبحلول وقت رحلة تشالنجر (المنحنى المتصل في الشكل أعلاه). فإن احتمال أي تلف حراري يزيد عن 0.6 حتى في الأجواء الحارة، ويصل إلى 1 في الرحلات التي تكون فيها درجة الحرارة 70 °F، ويجب ملاحظة أن درجة الحرارة الفعلية لمكوك الفضاء تشالنجر عند الإطلاق كانت 31 °F والتي يمكن وضعها في أعلى اليسار بالشكل (3.9).

الرسومات البيانية التشخيصية والإحصائيات التشخيصية:

Diagnostic Statistics and Plots

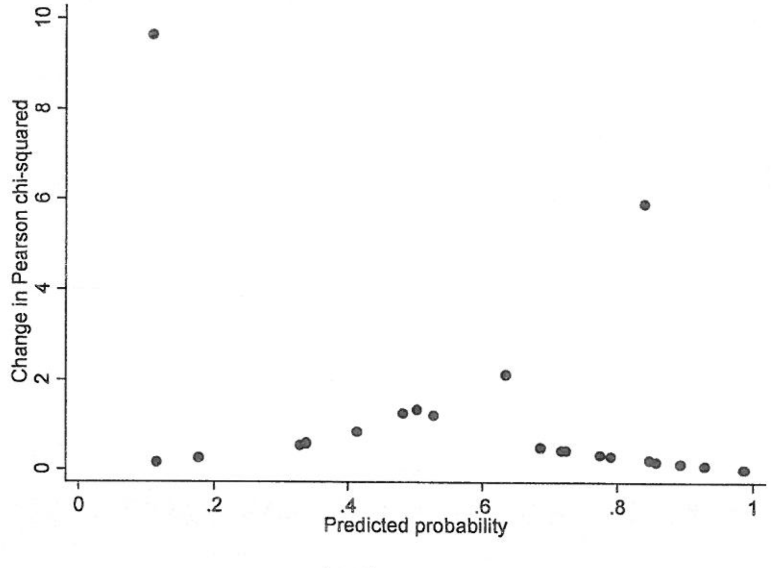
كما أشرنا سابقاً، فإن تأثير الانحدار اللوغاريتمي والإحصائيات التشخيصية التي يمكن الحصول عليها عن طريق استخدام الأمر **predict** لا تشير إلى مشاهدات فردية كما يحدث في حالة الإحصائيات التشخيصية

... OLS في الفصل (7). وبدلاً من ذلك، فإن التشخيصات اللوغاريتمية تشير إلى أنماط متغيرات x. وفي بيانات مكوك الفضاء فإن كل نمط لمتغير x لا نظير له، حيث لا توجد رحلتان تشتركان في نفس التاريخ *date* ودرجة الحرارة (وهذا طبيعي حيث لم يتم إطلاق رحلتين في نفس اليوم)، وقبل استخد... predict سوف نقوم بإعادة حساب آخر نموذج.

```
.quietly logistic any date temp
.predict Phat3
.label variable Phat3 "Pred...ability"
.predict dX2, dx2
.label variable dX2 "Change in Pearson chi-
 squared"
.predict dB, dbeta
.label variable dB "Influence"
.predict dD, ddeviance
.label variable ... "Change in deviance"
```

تشير دراسة Hosmer and Lemeshow (2000) ...ان الرسومات البيانية تساعد في قراءة الإحصائيات التشخيصية، ولإنشاء رسم بياني للتغير في χ^2 لبيرسون مع احتمال التلف الحراري (الشكل 4.9) نقوم بطباعة الأمر التالي:

```
.graph twoway scatter dX2 Phat3
```



الشكل (4.9)

هناك نمطان اثنان للمتغيرات *x* يتناسبان بشكل ضعيف وهذا واضح في أعلى اليمين واليسار بالشكل (4.9) حيث تبرز هاتان النقطتان بوضوح، يمكننا تحديد الرحلات التي لها قيم *dX2* مرتفعة بالعين المجردة من خلال إضافة توصيف لنقاط الانتشاربالرسم البياني. (في هذا المثال، سوف نستخدم رقم الرحلة *flight* للتوصيف)، في الشكل (5.9) سوف يتم توصيف الرحلات التي يكون فيها *dX2*> 2 فقط، وذلك بوضع هذه التوصيفات فوق شكل الانتشار (إذا قمنا بإضافة توصيفات لكل النقاط بالرسم البياني، فإننا لن نستطيع قراءة الجزء السفلي من البيانات التي تكون فيها النقاط متقاربة).

```
.graph twoway scatter dX2 Phat3
    ||  scatter  dX2  Phat3  if  dX2>  2,
    mlabel(flight)
    mlabsize(medsmall)
    || , legend(off)
```


الشكل (5.9)

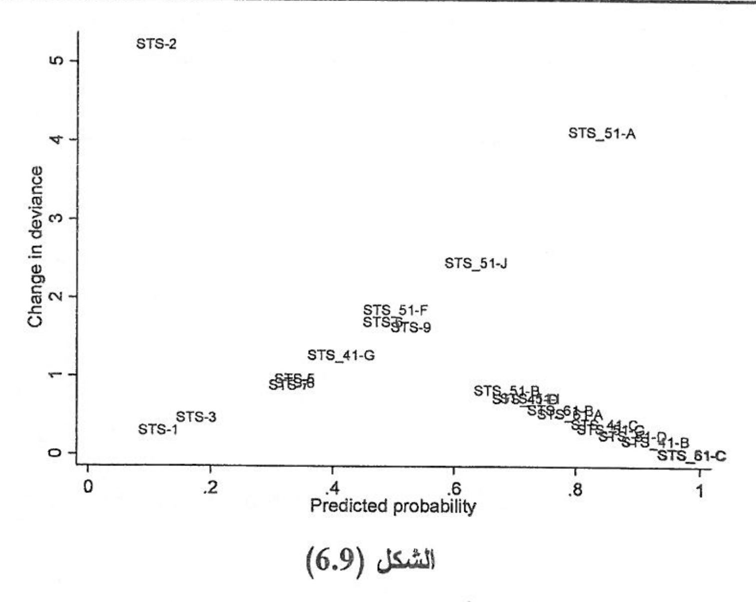
```
.list flight any date temp dX2 Phat3 if dX2> 2
```

	flight	any	date	temp	dX2	Phat3
2.	STS-2	1	12nov1981	70	9.630337	.1091805
4.	STS-4	.	27jun1982	80	.	.0407113
14.	STS_51-A	0	08nov1984	67	5.899742	.8400974
21.	STS_51-J	0	03oct1985	79	2.124642	.6350927
25.	STS_51-L	.	28jan1986	31	.	.9999012

الرحلة SAS 51-A لم تعان من أي تلف حراري بالرغم من تأخر تاريخ إطلاقها، وانخفاض درجة الحرارة (انظر الشكل 2.9)، النموذج يتوقع أن 0.84 احتمال وقوع تلف حراري لهذه الرحلة. كل النقاط التي تقع في الجانب الأيمن بالشكل (5.9) لم تعان من أي تلف حراري ($any = 0$) ولكن في أعلى اليسار ($any = 1$) الخاصة بالرحلة STS-2 عانت من تلف حراري بالرغم من أنها واحدة من الرحلات الأولى، وتم إطلاقها في جو معتدل، النموذج يتوقع بأن 0.109 احتمال وقوع تلف، وحيث إن ستاتا يعتبر القيم المفقودة أعلى قيم، فإنه يقوم بوضع القيمتين المفقودتين للرحلات بما فيها رحلة تشالنجر ضمن نطاق البيانات $dX2 > 2$.

نفس النتائج تم الحصول عليها من الرسم البياني للمتغير dD مع الاحتمال المتوقع كما هو معروض في الشكل (6.9). ومرة أخرى، فإن الرحلتين STS-2 (أعلى اليسار) وSTS 51-A (أعلى اليمين) تبرزان وتتناسبان بشكل ضعيف مع النموذج، الشكل (6.9) يوضح التباين في شكل الانتشار للنقاط التي تم توصيفها، وبدلاً من وضع رقم الرحلة بالقرب من النقاط كما فعلنا في الشكل (5.9) سوف نقوم بإخفاء النقاط نفسها عن طريق الخيار **msymbol(i)** ونضع توصيفات في مكان النقاط عن طريق الخيار **mlabposition(0)** لكل نقطة بيانات في الشكل (6.9).

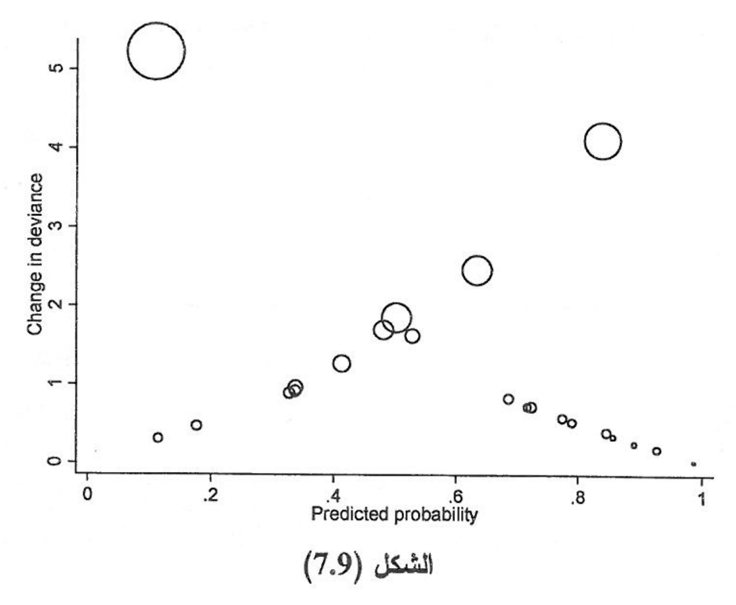
```
.graph twoway scatter dD Phat3, msymbol(i)
 mlabposition(0)mlabel(flight) mlabsize(small)
```



الشكل (6.9)

المتغير dB يقوم بقياس تأثير نمط متغير x في الانحدار اللوغـــاريتمي، الشكل (7.9) له نفس تصميم الشكل (6.9) ولكن يعرض النقاط على شـــكل دوائر حسب تأثيرها. الشكل يوضح بأن هناك مشاهدتين هما أسوأ المشاهدات وهما الأكثر تأثيراً.

```
.graph twoway scatter dD Phat3 [aweight = dB],
 msymbol(oh)
```



الشكل (7.9)

المشاهدات التي تتناسب بشكل ضعيف ولها تأثير كبير تستحق عناية خاصة لأنها تتعارض مع النمط العام للبيانات والنموذج المستخرج يتوقع اتجاهها المختلف، وبالطبع فإن استبعاد مثل تلك القيم المتطرفة يسمح لنا بالحصول على نموذج أكثر تناسباً مع بقية البيانات، ولكن هذا ما تعرضه الأشكال الدائرية بالرسم، ردة الفعل الأكثر قبولاً هو التحقق من سبب ظهور القيم المتطرفة فلماذا رحلة مكوك الفضاء STS-2 وليست الرحلة STS 51-A عانت من تلف أثناء الإطلاق؟ البحث عن إجابة لهذا السؤال قد تقود المحققين إلى النظر في المتغيرات السابقة.

الانحدار اللوغاريتمي مع الفئة المُرتبة y :

Logistic Regression with Ordered-Category y

الأمر **logit** والأمر **logistic** يتناسبان مع نماذج لها متغيرات، وهذه المتغيرات لها نوعان من المخرجات يتم ترميزهما 0 و 1، ونحن نحتاج إلى طرق أخرى للنماذج التي يأخذ فيها المتغير *y* أكثر من قيمتين. الاحتمالان المهمان هما الانحدار اللوغاريتمي المتعدد، أو المرتب:

mlogit الانحدار اللوغاريتمي المتعدد: الذي يكون فيه المتغير *y* له فئات متعددة ولكنها غير مرتبة مثل (1= ديمقراطي Democrat، 2= جمهوري Republican، 3= أخرى other).

ologit الانحدار اللوغاريتمي المرتب: الذي يكون فيه المتغير *y* متغيراً ترتيبياً (فئة مرتبة)، القيم الرقمية تمثل فئات ليست ذات أهمية إلا إذا كانت الأرقام الأعلى تشير إلى الكثرة، مثلاً فئات المتغير *y* قد تشير إلى (1= سيء poor، 2= مقبول fair، 3= ممتاز excellent).

قمنا سابقاً بتحديد المخرجات الثلاثة للمتغير الترتيبي *distress* في تفرعات للمتغير *any*، الأمر **logit** والأمر **logistic** يتطلبان متغيرات تابعة لها قيمتان {0,1}. ومن ناحية أخرى، فإن الأمر **ologit** تم تصميمه للمتغيرات الترتيبية التي لها أكثر من قيمتين، دعنا نقول بأن المتغير *distress* له عدة مخرجات هي 0= "لا شيء"، 1="1 أو 2"، 2="3 أو أكثر" وهي تشير إلى حوادث التلف بصواريخ الإطلاق.

الانحدار اللوغاريتمي المرتّب يشير إلى أن المتغير *date* والمتغير *temp* كليهما يؤثر في المتغير *distress* بنفس الإشارات (موجب لمتغير *date*، سالب لمتغير *temp*) كما شاهدنا سابقاً في التحليل اللوغاريتمي الثنائي:

```
.ologit distress date temp, nolog
```

```
Ordered logistic regression                       Number of obs   =         23
                                                  LR chi2(2)      =      12.32
                                                  Prob > chi2     =     0.0021
Log likelihood =  -18.79706                       Pseudo R2       =     0.2468
```

distress	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
date	.003286	.0012662	2.60	0.009	.0008043	.0057677
temp	-.1733752	.0834475	-2.08	0.038	-.3369293	-.0098212
/cut1	16.42813	9.554822			-2.298978	35.15524
/cut2	18.12227	9.722?02			-.933092	37.17763

اختبارات نسب الاحتمال أكثر دقة من اختبارات *z* المقاربة، أولاً يجب استخدام الأمر **estimates store** مع النتائج المحفوظة في الذاكرة من النموذج الكامل (مع متغيرين تنبؤيين) والذي تم حسابه سابقاً، يمكننا إعطاء هذا النموذج أي اسم توصيفي وليكن اسمه *date_temp*

```
.estimates store date_temp
```

ثم بعد ذلك القيام بإنشاء نموذج أكثر بساطة بدون المتغير *temp* وحفظ نتائج النموذج باسم *notemp* وإجراء اختبار نسبة الاحتمال لمعرفة ما إذا كان النموذج *notemp* يختلف بدرجة كبيرة عن النموذج الكامل *date_temp*

```
.quietly ologit distress date
.estimates store notemp
.lrtest notemp date_temp
```

```
Likelihood-ratio test                                 LR chi2(1)  =      6.12
(Assumption: notemp nested in date_temp)              Prob > chi2 =    0.0133
```

وتجدر الملاحظة أن مخرجات الأمر **lrtest** وافتراضه يشيران إلى أن نموذج *notemp* يتداخل مع نموذج *date_temp*، وهذا يعني أن المعلمات التي تم تقديرها في نموذج *notemp* هي عبارة عن مجموعة فرعية من معلمات

نموذج *date_temp* وكلا النموذجين تم تقديرهما من نفس مجموعة المشاهدات (والتي يمكن أن تكون معقدة عند احتواء البيانات على قيم مفقودة). اختبار الأرجحية العظمى يشير إلى أن نموذج *notemp* يتناسب بشكل ضعيف جداً، وحيث إن وجود المتغير *temp* كمتغير تنبؤي في نموذج *date_temp* هو الاختلاف الوحيد لذلك، فإن اختبار الأرجحية العظمى يوضح بأن مساهمة المتغير *temp* في النتائج هي مساهمة ذات أهمية كبيرة. وهناك خطوات أخرى مشابهة تؤكد بأن متغير *date* له تأثير ذو معنوية.

```
.quietly ologit distress temp
.estimates store nodate
.lrtest date_temp؛

Likelihood-ratio test                                 LR chi2(1)  =     10.33
(Assumption: nodate nested in date_temp)              Prob > chi2 =    0.0013
```

الأمر **estimates store** والأمر **lrtest** يعتبران أدوات مرنة لمقارنة نماذج الأرجحية العظمى. وللحصول على مزيد من التفاصيل عن هذه الأوامر والخيارات المصاحبة لها، قم بطباعة الأمر **help lrtest** أو الأمر **help estimates**.

النموذج اللوغاريتمي المرتب يقوم بتقدير نتيجة S لكل مشاهدة، حيث إن كل مشاهدة تكون دالة خطية للمتغير *date* والمتغير *temp*.

$$S = 0.003286date - 0.1733752temp$$

الاحتمالات المتوقعة تعتمد على قيمة S زائداً الاضطراب الموزع اللوغاريتمي u، الاقتراب من نقاط القطع المقدّرة (المعروضة في مخرجات الأمر **ologit** باستخدام cut1، cut2 ... الخ).

$$P(distress = \text{"none"}) = P(S+u \leq \text{cut1})$$
$$= (1 + \exp(-\text{cut1} + S))^{-1}$$
$$P(distress = \text{"1 or 2"}) = P(\text{cut1} < S + u \leq \text{cut2})$$
$$= (1+ \exp(-\text{cut2} + S))^{-1} - (1 + \exp(-\text{cut1} + S))^{-1}$$
$$P(distress = \text{"3 plus"}) = P(\text{cut2} < S + u)$$
$$= 1 - (1 + \exp(-\text{cut2} + S))^{-1}$$

بعد إجراء حساب الاحتمالات المتوقعة باستخدام الأمر **ologit** والأمر **predict** لكل فئة من فئات المتغير التابع، نقوم بإعطاء أسماء لكل الاحتمالات

التي تم حسابها بالأمر **predict**. فمثلاً الاسم *none* قد يشير إلى احتمال عدم وجود حادث تلف (الفئة الأولى لحوادث التلف *distress*)، والاسم *threeplus* يشير لثلاثة حوادث تلف أو أكثر (ثالث وآخر فئة لحوادث التلف *distress*):

```
.quietly ologit distress date temp
.predict none onetwo threeplus
```

هذا يقوم بإنشاء ثلاثة متغيرات جديدة:

```
.describe none onetwo threeplus
```

variable name	storage type	display format	value label	variable label
none	float	%9.0g		Pr(distress==0)
onetwo	float	%9.0g		Pr(distress==1)
threeplus	float	%9.0g		Pr(distress==2)

الاحتمالات المتوقعة لآخر رحلة لمكوك الفضاء تشالنجر – وهي المشاهدة رقم 25 في هذه البيانات – متذبذبة.

```
.list flight none onetwo threeplus if flight ==25
```

	flight	none	onetwo	threep~s
25.	STS_51-L	.0000754	.0003346	.99959

النموذج الذي قمنا بإنشائه يعتمد على تحليل 23 رحلة تسبق رحلة تحطّم مكوك تشالنجر، ويعطي احتمالاً ضئيلاً ($p = 0.000075$) بأن المكوك تشالنجر سوف لن يواجه تلفاً مصاحباً للانطلاق، واحتمال أكثر ضآلة بحادث أو اثنين ($p = 0.003$) ولكن الاحتمال الواضح والمؤكد ($p = 0.9996$) لوقوع ثلاثة حوادث تلف أو أكثر.

للحصول على مزيد من التفاصيل عن التقنيات المتعلقة بهذا النوع من التحليل، انظر دراسة Long (1997) ودراسة Hosmer and Lemeshow (2000) كما أن دليل المستخدم *the Reference Manual* يشرح تطبيقات برنامج ستاتا المتعلقة بهذا الموضوع. بالإضافة إلى ذلك، فإن دراسة Long and Freese (2006) تعرض نقاشاً أكثر تفصيلاً، وتشرح الخيارات المتوافرة عن القيام بهذا التحليل عن طريق

استخدام الملفات التنفيذية do-files، كما توضح بعض التفسيرات المفيدة وأوامر ما قبل التقدير مثل اختبارات Brant، ولتثبيت هذه الأوامر التنفيذيـــة المجانيـــة مـــن الإنترنت قم بطباعة الأمر **findit brant** ثم اتبع الرابط الذي يرشدك إلى مصادر تلك الملفات التنفيذية.

الانحدار اللوغاريتمي المتعدد : Multinomial Logistic Regression

إذا كانت فئات المتغير التابع غير ترتيبية، فإن الانحدار اللوغـــاريتمي المتعدد (يُسمى أيضاً الانحدار اللوغاريتمي متعـــدد التـــدريج) يـــوفر أدوات مناسبة؛ فإذا كان المتغير *y* له فئتان، فإن الأمر **mlogit** والأمر **ologit** كليهما يتناسب مع نفس النموذج كما يحدث مع الأمر **logistic**. إلا أن نموذج الأمر **mlogit** أكثر تعقيداً.

تظهر المتغيرات التابعة ذات الفئات المتعددة عادة في بيانات الدراسات الاستقصائية؛ فمثلاً بيانات استقصاء جرانيت توضح ذلك.

```
.use C:\data\Granite2011_6.dta, clear
.describe age sex educ party warmop2 warmice
```

variable name	storage type	display format	value label	variable label
age	byte	%9.0g	age	Age of respondent
sex	byte	%9.0g	sex2	Gender
educ	byte	%14.0g	educ	Highest degree completed
party	byte	%11.0g	party	Political party identification
warmop2	byte	%9.0g	yesno	Believe happening now/human
warmice	byte	%9.0g	warmice2	Arctic ice vs. 30 years ago

رأينا في الفصل (4) أن الاستطلاع يتضمن ثلاثة أسئلة رئيـــسة حـــول المناخ مثل المتعلقة بالمتغير *warmice*:

أي من العبارات الثلاث التالية تعتقد أنها أكثر دقة؟

خلال السنوات القليلة الماضية، جليد القطب الشمالي في آخـــر فـــصل الصيف ...

- *غطى أقل مساحة عن تلك التي كان يغطيهـــا فـــي الـــسنوات 30 الماضية.*

- *انخفض ولكنه عاد ليغطي نفس المناطق التي كان يغطيها في 30 سنة الماضية.*
- *يغطي مناطق أكثر من تلك التي كان يغطيها في 30 سنة الماضية.*

هذه البيانات تم اعتبارها بيانات دراسة استقصائية باستخدام الأمر **svyset** (انظر الفصل 4)، وهذه البيانات توضح معلومات عن الأوزان والمعاينة، والأوامر التي تُستخدم قبلها **svy** سوف تقوم تلقائياً بتطبيق هذه المعلومات، فمثلاً نسب الردود الموزونة للمتغير *warmice* تم الحصول عليها كما يلي:

```
.svy: tab warmice, percent
(running tabulate on estimation sample)

Number of strata   =         1           Number of obs      =       516
Number of PSUs     =       516           Population size    = 515.57392
                                         Design df          =       515

----------------------------
Arctic     |
ice vs.    |
30 years   |
ago        | percentages
-----------+----------------
      Less |      70.91
  Recovere |      10.43
      More |      6.916
     DK/NA |      11.75
           |
     Total |        100
----------------------------
  Key:  percentages  =  cell percentages
```

حوالي 71% من الذين شاركوا في الاستطلاع أجابوا بأن هناك انخفاضاً في جليد المنطقة الشمالية، وأن نسبة 12% فقط قالوا بأنهم لا يعرفون أو لم يقوموا بالإجابة.

المتغير الثاني الذي يهمنا يشير إلى ما إذا كان المشاركون في الدراسة يعتقدون بأن التغير المناخي يحدث في الوقت الحاضر وسببه الأنشطة البشرية (*warmop2*)، حوالي 55% يعتقدون أن هذه العبارة صحيحة.

```
.svy: tab warmop2, percent
```

```
(running tabulate on estimation sample)

Number of strata   =          1              Number of obs      =        516
Number of PSUs     =        516              Population size    = 515.57392
                                             Design df          =        515
```

Believe happening now/human	percentages
No	45.11
Yes	54.89
Total	100

Key: percentages = cell percentages

إجابات السؤال المتعلق بالمتغير *warmice* ترتبط مع الاعتقـاد بوجـود التغير المناخي، كما نرى في الجدول ذي الاتجاهين أدناه مع نسب مئوية بناءً على متغير الصف بالجدول وهو *warmop2*، نسبة الإجابات الدقيقة الخاصـة بالمتغير *warmice* كانت 83% وهم الذين يعتقدون أن الإنسان يتـسبب فـي تغييرات في المناخ، ولكن 56% فقط لا يعتقدون ذلك، أما الإجابة المخالفـة للواقع، وهي أن الجليد في آخر فصل صيف بالمنطقة القطبية الشمالية غطى نفس المناطق التي كان يغطيها منذ 30 سنة مضت. هذه الإجابة كانت أربع مرات أكثر شيوعاً بين الذين لا يعتقدون أن الإنسان يتسبب في تغييرات في المناخ، هذه الاختلافات ذات معنوية إحصائية عند ($p \approx 0.0000$).

.svy: tab *warmop2 warmice*, row percent

```
(running tabulate on estimation sample)

Number of strata   =          1              Number of obs      =        516
Number of PSUs     =        516              Population size    = 515.57392
                                             Design df          =        515
```

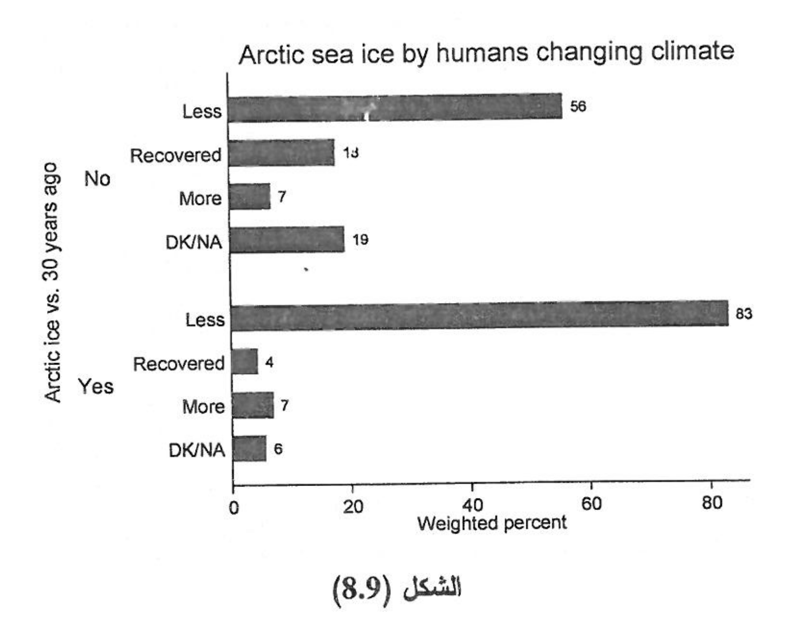
Believe happening now/human	Arctic ice vs. 30 years ago				
	Less	Recovere	More	DK/NA	Total
No	56.07	17.81	6.951	19.17	100
Yes	83.1	4.354	6.887	5.654	100
Total	70.91	10.43	6.916	11.75	100

Key: row percentages

```
Pearson:
  Uncorrected   chi2(3)           =   55.2306
  Design-based  F(3.00, 1544.45)=   14.6772      P = 0.0000
```

الفصل (4) يعرض الأمر **catplot** الذي يساعد في إنشاء رسومات بيانية للمتغيرات النوعية، الأمر **catplot** لا يتم تثبيته افتراضياً مع برنامج ستاتا، ولكن يجب تحديده وتحميله عن طريق طباعة الأمر **findit catplot**، باستخدام الأمر **catplot** يمكننا إنشاء أعمدة بيانية تتعلق بالجدول الثنائي أعلاه، عند اعتبار المتغيرالموزون الاحتمالي *censuswt* كأوزان تحليلية (**[aw=censuswt]**) فإن النتائج في الأعمدة البيانية سوف تتشابه مع النسب الناتجة من الأمر **svy: tab**.

```
.catplot hbar warmice [aw=censuswt],
 over(warmop2)percent(warmop2)
 blabel(bar, format(%2.0f)) ytitle("Weighted
 percent")
 title("Arctic sea ice by humans changing
 climate")
```



الشكل (8.9)

الشكل (8.9) يشبه جدول النسب، واختبار F، حيث يعرض العلاقة بين الاعتقادات والحقائق حول المناخ. يمكنك فحص ما إذا كانت هناك أنماط متشابهة ظهرت مع السؤالين الرئيسين في استطلاع الرأي، المتغير *warmco2* (الاتجاهات في CO_2) والمتغير *warmgre* (تأثير الصوبات الزجاجية)، التفسير

التقليدي لمثل هذه الأنماط هو أن المعرفة توفر معلومات لمن لديهم اعتقادات حول المناخ، وفي هذه الحالة معرفة حقائق محددة حول المناخ توضح ما إذا كان الناس يعتقدون أن الإنسان يتسبب في تغير المناخ. وعموماً فإن بحوث العلوم الاجتماعية التي تم إجراؤها حديثاً وجدت دليلاً على علاقة سببية في الاتجاه المعاكس، بعض الناس يقبلون حقيقة محددة أو حقائق خاطئة، لأنها تتناسب مع اعتقاداتهم عموماً.

وللاستمرار مع هذه الفرضيات (اصطلاح "الاستيعاب المتحيّز") يمكننا تحليل الردود على المتغير *warmice* باعتباره متغيرًا تابعاً. المتغيرات التنبؤية المحتملة تتضمن العمر والجنس والتعليم ووجهة النظر السياسية. وعادة فإن هذه المتغيرات التنبؤية ترتبط مع وجهات النظر المتعلقة بالبيئة. في بيانات الدراسة الاستقصائية الموجودة لدينا متغير العمر *age* يحتوي على سنوات، ومتغير الجنس *sex* يحتوي على الرقم 0 للذكور والرقم 1 للإناث، ومتغير التعليم *educ* يتراوح ما بين الرقم 1 للتعليم الثانوي أو أقل، إلى الرقم 4 للدراسات العليا، أما متغير الانتماء السياسي *party* يتضمن الرقم 1 للديمقراطيين والرقم 2 للمستقلين والرقم 3 للجمهوريين.

```
.describe age sex educ party warmop2 warmice

              storage  display    value
variable name   type   format     label      variable label
------------------------------------------------------------------------------
age             byte   %9.0g      age        Age of respondent
sex             byte   %9.0g      sex2       Gender
educ            byte   %14.0g     educ       Highest degree completed
party           byte   %11.0g     party      Political party identification
warmop2         byte   %9.0g      yesno      Believe happening now/human
warmice         byte   %9.0g      warmice2   Arctic ice vs. 30 years ago
```

كيف تؤثر هذه العوامل على ردود المشاركين حول الأسئلة المتعلقة بالمتغير *warmice*؟ هل الاعتقادات السائدة حول التغير المناخي يمكنها أن تساهم في توقع الإجابات إذا ما قمنا بالتحكم في التعليم والانتماء السياسي؟ نتائج الأمر **mlogit** أدناه تعطي بعض الإجابات عن هذه الأسئلة.

```
.svy: mlogit warmice age sex educ party
 warmop2, rrr base(1)
```

```
(running mlogit on estimation sample)

Survey: Multinomial logistic regression

Number of strata   =          1          Number of obs      =        486
Number of PSUs     =        486          Population size    =  485.77734
                                         Design df          =        485
                                         F(  15,     471)   =       4.54
                                         Prob > F           =     0.0000
```

warmice	RRR	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
Less	(base outcome)					
Recovered						
age	1.001732	.0110398	0.16	0.875	.9802738	1.023661
sex	.6975992	.2518093	-1.00	0.319	.3432281	1.417846
educ	.8860304	.1491035	-0.72	0.472	.6365725	1.233245
party	1.718036	.4143614	2.24	0.025	1.069604	2.759569
warmop2	.239992	.1098955	-3.12	0.002	.097599	.590131
_cons	.1196324	.1170444	-2.17	0.030	.0174976	.8179363
More						
age	1.023417	.0147491	1.61	0.109	.9948431	1.052811
sex	.5854667	.2578116	-1.22	0.225	.2464541	1.390812
educ	.5378248	.0936788	-3.56	0.000	.3819503	.7573119
party	1.169189	.3220132	0.57	0.571	.6805561	2.008656
warmop2	1.270082	.6225808	0.49	0.626	.4847726	3.327558
_cons	.0833092	.0846524	-2.45	0.015	.0113137	.6134542
DK_NA						
age	.9866127	.0109802	-1.21	0.226	.9652723	1.008425
sex	1.253388	.4430697	0.64	0.523	.625799	2.51036
educ	.8338215	.1369808	-1.11	0.269	.6037919	1.151487
party	1.707791	.3624779	2.52	0.012	1.125423	2.591516
warmop2	.2443751	.1004212	-3.43	0.001	.1089927	.5479193
_cons	.2678258	.2778029	-1.27	0.205	.0348926	2.055758

هذا المثال يستخدم أوزان بيانات الدراسة الاستقصائية. تركيبة الأمر أعلاه يمكن أن تتشابه (ولكن بدون استخدام **svy** قبل الأمر) إذا لم نستخدم بيانات دراسة استقصائية، الخيار **base(1)** يحدد الفئة 1 ("مناطق أقل" = *warmice*) التي يُفترض أن تكون نتيجة أساسية للمقارنة، وبذلك فإن الجدول المعروض أعلاه يعرض المتغيرات التنبؤية لثلاث إجابات خاطئة مختلفة، أما

الخيار **rrr** يطلب من الأمر **mlogit** أن يعرض نسب المخاطرة النسبية، والتي تشبه نسب الاحتمال التي يمكن الحصول عليها بالأمر **logistic**.

وبصفة عامة، فإن نسب المخاطرة النسبية للنتيجة j الخاصة بالمتغير y والمتغير التنبؤي x_k تساوي نسبة الاحتمال المتوقعة لصالح $y = j$ (مقارنة مع "النتيجة الأساسية" = y) المضروبة في 1□ وحدة نقص واحدة في x_k مع ثبات العوامل الأخرى. بعبارة أخرى، فإن نسبة المخاطرة النسبية rrr_{jk} عبارة عن المضروب في حالة أن جميع متغيرات x عدا x_k تظل بنفس قيمتها.

$$\text{rrr}_{jk} \times \frac{P(y=j \mid x_k)}{P(y=\text{base} \mid x_k)} = \frac{P(y=j \mid x_k+1)}{P(y=\text{base} \mid x_k+1)}$$

نسب المخاطرة النسبية في المثال أعلاه تشرح التأثير المضاعف لزيادة وحدة واحدة في كل متغير تنبؤي على احتمال اختيار إجابة معينة خاصة بالمتغير *warmice* بدلاً من فئة أساسية (إجابة صحيحة) وهي الإجابة "مناطق أقل".

إننا نرى أن الجمهوريين لهم تأثير ذو معنوية إحصائية ($p = 0.025$) أكثر احتمالاً ليعتقدوا بأن الجليد عاد من جديد، بينما الذين يعتقدون أن أسباباً بشرية وراء التغير المناخي هم أيضاً لهم تأثير ذو معنوية إحصائية ($p = 0.002$) أقل احتمالاً ليعتقدوا ما يعتقده الجمهوريون. وفي حال ثبات العوامل الأخرى، فإن احتمالات أن أحد الجمهوريين ستكون إجابته أن الجليد عاد إلى مستوياته السابقة (بدلاً من تغطية مناطق أقل) هي 72% أعلى (مضروبة في 1.72) بالمقارنة مع المستقلين، وأعلى بنسبة 196% (مضروبة في $2.96=1.72^2$) من الديمقراطيين، أما الذين يعتقدون أن التغير المناخي يحدث حالياً نتيجة الإنسان هم 76% أقل احتمالاً (مضروبة في 0.24) لتكون إجابتهم أن الجليد عاد إلى مستوياته السابقة بدلاً من انخفاضه.

الجزءان الثاني والثالث من الجدول يعرضان مخرجات الأمر **mlogit** التي توضح نسب المخاطرة النسبية لصالح كل من الردود الأخرى للمتغير

warmice بالمقارنة مع من "يغطي مناطق أقل"، الإجابة التي تُقرّ بأن جليد القطب الشمالي يغطي مناطق أكثر من تلك التي كان يغطيها خلال 30 سنة ماضية هي إجابة مفضّلة من قبل المشاركين الأقل تعليماً، واحتمالات الحصول على هذه الإجابة انخفض بنسبة 46% (مضروبة في 0.54) مع كل زيادة بمقدار وحدة واحدة في متغير *educ* مع ثبات العوامل الأخرى. وعليه، فإن هذه الإجابة "عودة الجليد ليغطي مناطق أكثر" لها متغيرات تنبؤية تتعلق بالمعتقدات أو الانتماء السياسي، بينما الإجابة بأن "الجليد يغطي مناطق أكثر من ذي قبل" تبدو أنها تعكس عدم المعرفة لدى المشارك في الدراسة، أما إجابة "لا أعرف" أو "عدم الحصول على إجابة" (DK/NA) لها علاقة بمتغيرات تنبؤية تتعلق بالمعتقدات أو الانتماء السياسي، وربما تشير إلى نوع من الرفض للسؤال.

الأمر **margins** والأمر **marginsplot** يمكنهما تمثيل النتائج بيانياً، الأوامر أدناه تقوم بإنشاء رسم بياني تقريبي (الشكل 9.9) يعرض الاحتمالات المتوقعة بأن تكون الإجابة عن المتغير *warmice* هي "مناطق أقل" “less area” كدالة للاعتقادات عن المناخ (*warmop2*) والانتماء السياسي بناءً على النموذج السابق للأمر **mlogit**، وعند إضافة الخيار **predict(outcome(1))** للأمر **margins** يجعلنا نركز على نتيجة المتغير التابع "مناطق أقل" “less area”

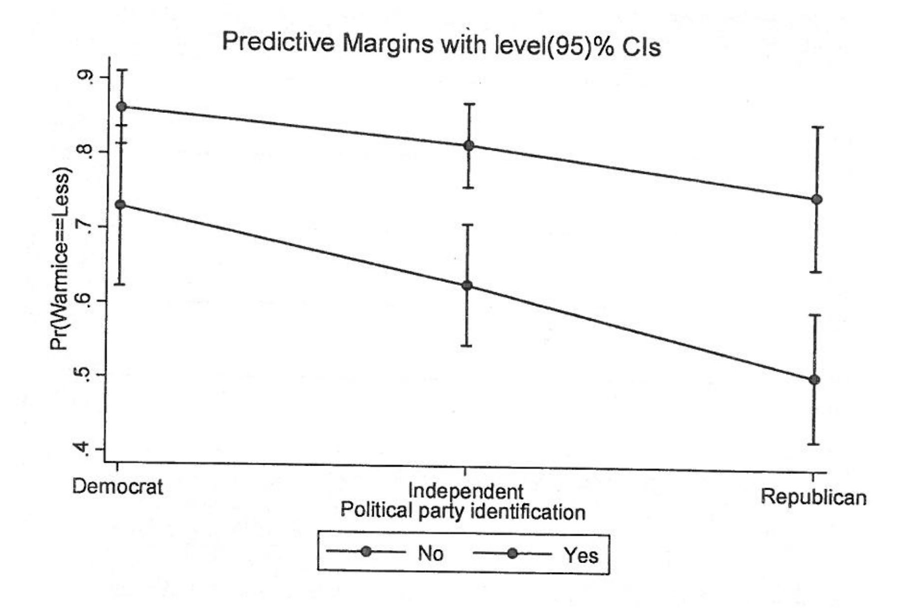
```
.margins, at(party = (1 2 3) warmop2 = (1 0))
 vsquish predict(outcome(1))
```

	Margin	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
_at						
1	.8607879	.0250921	34.31	0.000	.8116084	.9099675
2	.7289992	.0544082	13.40	0.000	.622361	.8356373
3	.8143384	.0286563	28.42	0.000	.7581732	.8705037
4	.626633	.0416184	15.06	0.000	.5450624	.7082036
5	.7500431	.0495879	15.13	0.000	.6528526	.8472336
6	.5072642	.0445321	11.39	0.000	.419983	.5945455

```
Predictive margins
.Model VCE      : Linearized                         Number of obs

Expression      : Pr(warmice==Less), predict(outcome(1))
1._at           : party           =          1
                  warmop2         =          1
2._at           : party           =          1
                  warmop2         =          0
3._at           : party           =          2
                  warmop2         =          1
4._at           : party           =          2
                  warmop2         =          0
5._at           : party           =          3
                  warmop2         =          1
6._at           : party           =          3
                  warmop2         =          0
```

. marginsplot



الشكل (9.9)

ولإنشاء رسم بياني للاحتمال المتوقع للحصول علـى الإجابـة الثانيـة للمتغير *warmice* تكون "عودة الجليـد" "recovered" نقـوم بتكـرار الأمـر **margins** ولكن مع الخيار **predict(outcome(2))**، الـشكل (10.9) يعـرض النتائج.

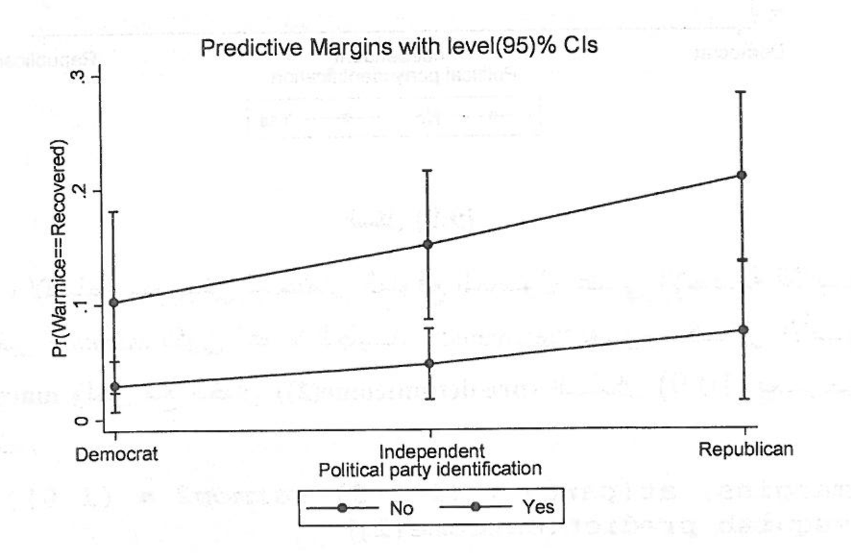
```
.margins, at(party = (1 2 3) warmop2 = (1 0))
 vsquish predict(outcome(2))
```

```
Predictive margins                              Number of obs     =        486
Model VCE     : Linearized

Expression    : Pr(warmice==Recovered), predict(outcome(2))
1._at         : party           =           1
                warmop2         =           1
2._at         : party           =           1
                warmop2         =           0
3._at         : party           =           2
                warmop2         =           1
4._at         : party           =           2
                warmop2         =           0
5._at         : party           =           3
                warmop2         =           1
6._at         : party           =           3
                warmop2         =           0
```

	Margin	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
_at						
1	.0290269	.0112526	2.58	0.010	.0069723	.0510815
2	.1023611	.0404401	2.53	0.011	.0231001	.1816222
3	.0470891	.0156777	3.00	0.003	.0163614	.0778168
4	.1507864	.0329816	4.57	0.000	.0861437	.2154292
5	.0743527	.0306377	2.43	0.015	.014304	.1344015
6	.2091542	.0372832	5.61	0.000	.1360805	.282228

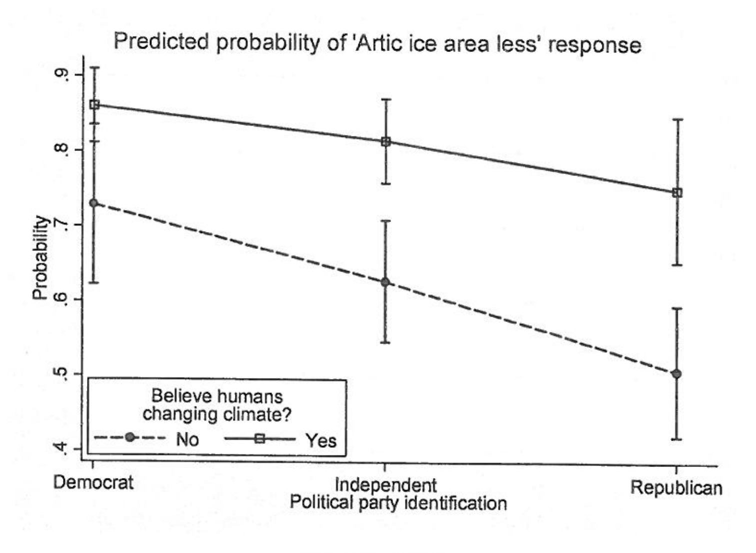
.marginsplot



الشكل (10.9)

الشكل (11.9) هو عبارة عن نسخة مطوّرة من الشكل (9.9) وهو يوضح استخدام بعض خيارات الأمر **marginsplot**، ونبدأ بالخيار **quietly** ونكرر استخدام الأمر **margins** للإجابة 1 (مناطق أقل) ثم نقوم بطباعة الأمر **marginsplot** مع خيارات أخرى للتحكم في تفاصيل التوصيفات بالرسم ومربع الشرح والخطوط، تم توسيع قياسات المحور الأفقي من 1 إلى 3.1 بدلاً من الوضع الافتراضي وهو من 1 إلى 3 وذلك حتى نستطيع وضع التوصيف "جمهوري" "Republican" داخل الرسم البياني بالأسفل جهة اليمين.

```
.quietly margins, at(party = (1 2 3) warmop2 =
 (1 0))   predict(outcome(1))
.marginsplot,      legend(position(7)     ring(0)
 rows(1)
 title("Believe  humans"  "changing  climate?",
 size(medsmall)))
 xscale(range(1 3.1)) ytitle("Probability")
 plot1opts(lpattern(dash)       lwidth(medthick)
 msymbol(O))
 plot2opts(lpattern(solid)      lwidth(medthick)
 msymbol(Sh))
 title("Predicted  probability  of  'Artic  ice
 area less'
 response")
```



الشكل (11.9)

الإسناد المتعدد للقيم المفقودة – مثال الانحدار اللوغاريتمي :
Multiple Imputation of Missing Values – Logit Regression Example

الفصل (8) عرض طرق الإسناد المتعدد للقيم المفقودة، مستخدماً مثـالاً عن الانحدار، طرق الإسناد المتعدد تعمل مع الأنواع الأخرى من الانحـدار بما فيها الانحدار اللوغاريتمي الذي تم شرحه في هذا الفصل، ولتوضيح ذلك سوف نعود لاستخدام بيانات استطلاع جرانيت، ومؤشر الاعتقـاد بوجـود التغير المناخي *warmop2*. الجزء السابق اختبر العمر والجـنس ومـستوى التعليم والانتماء السياسي كمتغيرات تنبؤية محتملة للإجابة عن سؤال المعرفة بالمناخ *warmice*. هذه الخصائص الأربع عادة يُعتقد بأن لهـا علاقـة فـي البحوث المتعلقة بالتأثيرات الاجتماعية للمشاكل البيئية، لذلك فمن المعقـول الاعتقاد بأن أحدها أو أكثر سوف تكون لها علاقة بالمتغير *warmop2*، وهل يُفترض أن نأخذ في الاعتبار دخل رب الأسرة والخصائص المهمة الأخرى كمتغير تنبؤي محتمل؟ إحدى المشاكل المتعلقـة بـدخل رب الأسـرة فـي استطلاع جرانيت هي أنها تحتوي على عدد كبير من القـيم المفقـودة، لأن الكثير من المشاركين لا يميلون للإجابة عن هذا السؤال.

عشر متغيرات من ملـف البيانـات *Granite2011_06.dta* سـوف يـتم استخدامها في هذا التحليل، أربعة منهـا (*employ, ownrent, married, yrslive*) نظرياً ليست لها أهمية عند الحديث عن الاعتقاد بوجود التغيـر المنـاخي، ولكنها ربما تساعد في إسناد القيم المفقودة لمتغير دخل رب الأسرة *income*.

```
.use C:\data\Granite2011_6.dta, clear
.describe warmop2 age sex educ party income
 employ ownrent married yrslive

              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------
warmop2         byte   %9.0g       yesno      Believe happening now/human
age             byte   %9.0g       age        Age of respondent
sex             byte   %9.0g       sex2       Gender
educ            byte   %14.0g      educ       Highest degree completed
party           byte   %11.0g      party      Political party identification
income          byte   %9.0g       income3    Household income 2009
employ          byte   %13.0g      employ     Employment status
ownrent         byte   %13.0g      ownrent    Own or rent home
married         byte   %9.0g       yesno      Respondent married
yrslive         byte   %8.0g       yrslive    Years lived in NH
```

```
.misstable summarize warmop2 age sex educ party
 income employ ownrent married yrslive
```

Variable	Obs=.	Obs>.	Obs<.	Obs<. Unique values	Min	Max
age	23		493	74	18	94
educ		5	511	4	1	4
party		13	503	3	1	3
income	171		345	7	1	7
employ		16	500	8	1	8
ownrent		20	496	2	0	1
yrslive		12	504	4	1	4

بالرغم من أننا قمنا بإدراج المتغيرات *warmop2, sex, married* ضمن الأمر **misstable** فإن ستاتا كشف أن هذه المتغيرات لا تحتوي على قيم مفقودة، ولم يتم عرضها ضمن المخرجات. ومن ناحية أخرى، فإنه لدينا 171 قيمة مفقودة من أصل 516 قيمة بمتغير *income*. وإذا قمنا بحساب انحدار المتغير الثنائي *warmop2* على المتغير *income* مع المتغيرات التي يمكن أن يكون لها علاقة، فإن التقديرات تتضمن 340 مشاهدة فقط.

```
.svy: logit warmop2 age sex educ party income
```

(running logit on estimation sample)

Survey: Logistic regression

Number of strata = 1		Number of obs	= 340
Number of PSUs = 340		Population size	= 336.84437
		Design df	= 339
		F(5, 335)	= 13.47
		Prob > F	= 0.0000

warmop2	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	-.014292	.0092398	-1.55	0.123	-.0324666	.0038826
sex	.4762698	.2829395	1.68	0.093	-.0802685	1.032808
educ	.2508435	.1525283	1.64	0.101	-.0491775	.5508646
party	-1.176907	.1547857	-7.60	0.000	-1.481369	-.8724461
income	.0366035	.0768294	0.48	0.634	-.114519	.1877259
_cons	2.252896	.7778417	2.90	0.004	.7228924	3.7829

المتغير الوحيد الذي له تأثير ذو المعنوية الإحصائية هو الانتماء السياسي. هل سوف نحصل على نفس النتيجة إذا ما قمنا بإجراء هذا التحليل من جديد بدون استبعاد أي بيانات جانباً؟ الإسناد المتعدد يُعتبر وسيلة للإجابة على هذا السؤال.

نقوم بأول خطوة في عملية الإسناد وهي استبعاد 42 مشاهدة لها قيم مفقودة في أي من المتغيرات ذات العلاقة باستثناء متغير *income*، وبعد القيام بذلك، سوف تكون لدينا بيانات تحتوي على 516 − 42 = 474 مشاهدة تتضمن 137 مشاهدة هي عبارة عن قيم مفقودة خاصة بالمتغير *income*.

```
.keep if !missing(warmop2, age, sex, educ,
 party, employ,ownrent, married, yrslive)
 (42observations deleted)
.misstable summarize warmop2 age sex educ party
 income employ ownrent married yrslive
```

			Obs<.			
Variable	Obs=.	Obs>.	Obs<.	Unique values	Min	Max
income	137		337	7	1	7

في الخطوة التالية، سوف نقوم بتحديد تنسيق لبيانات الإسناد المتعدد باستخدام الخيار**mlog** وهو اختيار لكفاءة الذاكرة. المتغير *income* سوف يتم تسجيله باستخدام الخيار **imputed** وهذا يعني أننا سوف نحاول أن نقوم بتعويض القيم المفقودة للمتغير *income*، أما المتغيرات الأخرى فسوف يتم تسجيلها كمتغيرات عادية **regular** ولن يتم إسنادها.

```
.mi set mlong
.mi register imputed income
 (137m=0 obs. now marked as incomplete)
.mi register regular warmop2 sex educ party
 employ ownrent married yrslive
```

137 مشاهدة قيمها مفقودة بالمتغير *income* والتي تم توقعها باستخدام الانحدار مع المتغيرات *employ, ownrent, married, yrslive*، تم إنشاء 50 قيمة

إسناد لكل القيم المتوقعة 137 زائداً تذبذباً عشوائياً، عمليات الإسناد هذه تـم تجميعها لتقدير نموذج انحدار لوغاريتمي جديد.

.mi impute regress *incomeemploy ownrent married yrslive,* add(50) rseed(12345)

```
Univariate imputation                    Imputations =      50
Linear regression                              added =      50
Imputed: m=1 through m=50                    updated =       0
```

Variable	Observations per m			
	Complete	Incomplete	Imputed	Total
income	337	137	137	474

```
(complete + incomplete = total; imputed is the minimum across m
 of the number of filled-in observations.)
```

.mi estimate: svy: logit *warmop2 age sex educ party income*

```
Multiple-imputation estimates              Imputations       =        50
Survey: Logistic regression                Number of obs     =       474

Number of strata   =          1            Population size   = 472.04182
Number of PSUs     =        474
                                           Average RVI       =    0.0298
                                           Largest FMI       =    0.1511
                                           Complete DF       =       473
DF adjustment:   Small sample              DF:     min       =    338.55
                                                   avg       =    443.42
                                                   max       =    470.18
Model F test:       Equal FMI              F(   5,  469.8)   =     16.88
Within VCE type:   Linearized              Prob > F          =    0.0000
```

warmop2	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	-.0188856	.0075324	-2.51	0.013	-.033687	-.0040842
sex	.4338802	.2372722	1.83	0.068	-.0323649	.9001254
educ	.2489546	.1262441	1.97	0.049	.0008665	.4970427
party	-1.154414	.1318129	-8.76	0.000	-1.41343	-.8953988
income	.0134369	.0681875	0.20	0.844	-.1206877	.1475614
_cons	2.669003	.6581604	4.06	0.000	1.375593	3.962413

بعد إجراء عمليات الإسناد المتعدد، فإن مُعَامِل المتغير *party* يظل قريباً من سابقه (1.15− بالمقارنة مع السابق وهو 1.18−) ولايـزال تـأثيره ذا معنوية إحصائية. النتائج الأخرى تُظهر تغيرًا أكبر، وهذا التغيـر يتـضمن تغيراً في المُعَامِلات ولكن بصفة عامة تتشابه مع الأخطاء المعياريـة ممـا يعكس أن التقديرات أصبحت أكثر دقة بعد إجراء عمليات الإسناد للبيانـات، من خلال هذه التعديلات فإن مُعَامِلات المتغير *age* (سالب) والمتغيـر *educ* (موجب) الآن أصبحت ذات معنوية إحصائية، وهذا يـشبه النتـائج التـي أظهرتها دراسات سابقة حول الاعتقاد بوجود التغير المناخي. ومـن ناحيـة أخرى، فإن المتغير *income* يبدو أقل تأثيراً قبل وبعد إجراء عملية الإسـناد. هذه النتيجة تدعم وجهة النظر التي تقترح استبعاد متغير *income* من النموذج النهائي، والتركيز على المتغيرات التنبؤية الأكثر أهمية والأقل إشكالاً.

الفصل العاشر

نماذج عد الأحداث والبقاء

Survival and Event-Count Models

يعرض هذا الفصل، نماذج لتحليل بيانات الحدث وتحليل البقاء. تحليل البقاء يتضمن عدة تقنيات ذات صلة تركز على أزمنة وقوع الحدث، وبالرغم من أن الحدث يمكن أن يكون حسناً أو سيئاً، فإننا سوف نتفق على أن نشير إلى أن الحدث باسم "الفشل". والوقت اللازم للفشل يسمى "زمن البقاء". تحليل البقاء مهم جداً في البحوث الطبية الحيوية، ولكن يمكن تطبيقه بشكل متساو على الحقول العلمية الأخرى من الهندسة إلى العلوم الاجتماعية. فمثلاً عند صياغة نموذج للزمن اللازم للشخص العاطل عن العمل للحصول على وظيفة، أو الشخص العازب حتى يتزوج، فإن برنامج ستاتا يوفر عددًا كبيرًا من طرق تحليل البقاء. وهذا الفصل سوف يشرح جزءًا بسيطاً منها.

وسوف نلقي نظرة سريعة على انحدار بواسون ومكوناته. هذه الطرق لا تركز فقط على أزمنة البقاء، ولكن تركز أيضاً على معدلات أو أعداد الأحداث خلال فاصل زمني محدد. طرق عد الأحداث تتضمن انحدار بواسون، والانحدار الثنائي السالب. مثل هذه النماذج يمكن صياغتها باستخدام أوامر محددة أو باستخدام المدخل الأكثر شمولاً، وهو النموذج الخطي العام generalized linear modeling (GLM).

يرجى الاطلاع على دليل المستخدم *Survival Analysis and Epidemiological Tables Reference Manual* للحصول على مزيد من المعلومات عن إمكانات برنامج ستاتا. أو قم بطباعة الأمر **help st** للاطلاع على نظرة عامة على شبكة الإنترنت عن هذا التحليل. كما أن دراسة Selvin (2004) تقوم بتوضيح

تحليل البقاء وانحدار بواسون. وقد قام مؤلف هذا الكتاب باستعارة (بعد الحصول على إذن بذلك) بعض الأمثلة من دراسة Selvin. الشرح المفيد الآخر عن تحليل البقاء يمكن الحصول عليه من كتاب خاص عن برنامج ستاتا تم إعداده من قبل .Vleves et al (2010)، وهناك فصل تم إعداده من قبل Rosner (1995)، كما تم إعداد شرح متكامل من قبل Hosmer, Lemeshow and May (2008) وLee (1992)، وتم شرح النماذج الخطية العامة من قبل McCullagh and Nelder (1989)، وفي كتاب Long (1997) هناك فصل حول نماذج الانحدار لعد البيانات (يشمل بواسون وذا الحدين السالب) وأيضاً بعض الأمثلة عن النماذج الخطية العامة. ويمكن الحصول على شرح مفصّل عن النماذج الخطية العامة في كتاب Hardin and Hilbe (2012).

مجموعات قوائم ستاتا التي لها علاقة بهذا الفصل تتضمن:

Statistics > Survival analysis
Graphics > Survival analysis graphs
Statistics > Count outcomes
Statistics > Generalized linear models

وتجدر الإشارة بأن الجداول الوبائية لم يتم تناولها في هذا الفصل، وللحصول على معلومات عنها قم بطباعة الأمر **help epitab** أو قم بالاطلاع على قوائم ستاتا:

Statistics > Epidemiology and related

أمثلة عن الأوامر : Example Commands

أغلب أوامر تحليل البقاء ببرنامج ستاتا (st*) تتطلب أن تكون البيانات قد تم تحديدها مسبقاً كزمن بقاء باستخدام الأمر **stset**، ويجب استخدام الأمر **stset** مرة واحدة، وبعد ذلك سوف يتم حفظ البيانات.

```
.stset timevar, failure(failvar)
```

يقوم بتحديد سجل مفرد ليمثل بيانات زمن بقاء المتغير *timevar* يُشير إلى الزمن الذي مضى قبل وقوع حدث معين (يسمى "الفشل") أو الفترة التي انتهت فيها المشاهدة ("المراقبة"). أما المتغير *failvar* يشير إلى ما إذا كان

الفشل (*failvar* = 1) أو المراقبة (*failvar* = 0) عند الزمن *timevar*. البيانات تحتوي فقط على سجل واحد لكل حدث، ويجب تحديدها باستخدام الأمر **stset** قبل استخدام أوامر **st*** في أي حسابات. وإذا قمنا بحفظ البيانات، فإن تحديدات الأمر **stset** سوف يتم حفظها أيضاً. الأمر **stset** يقوم بإنشاء متغيرات جديدة اسمها *_st, _d, _t, _t0* وهي عبارة عن ترميز للمعلومات اللازمة لاستخدام أي أمر من أوامر **st***.

```
.stset timevar, failure(failvar) id(patient)
 enter(time start)
```

يقوم بتحديد بيانات أزمنة البقاء لسجلات متعددة، وفي هذا المثال، فإن المتغير *timevar* يشير إلى الوقت الذي مضى قبل حدوث الفشل أو المراقبة. المتغير *failvar* يشير إلى ما إذا كان الفشل (1) أو المراقبة (0) حدثت عند هذا الوقت. المتغير *patient* عبارة عن رقم محدد. نفس الحالة يمكن أن تساهم في أكثر من سجل واحد في البيانات، ولكنها دائماً لها نفس الرقم المحدد. المتغير *start* يقوم بتسجيل زمن مراقبة كل مشاهدة.

```
.stdescribe
```

يشرح بيانات زمن البقاء، ويضع قائمة بالتعريفات والخصائص الأخرى للبيانات التي قام بإنشائها الأمر **stset**.

```
.stsum
```

يقوم بإنشاء إحصائيات مختصرة تتضمن: الوقت الكلي عند الخطر، معدل الوقوع، عدد المجموعات، ونسب أزمنة البقاء.

```
.ctset time nfail ncensor nenter, by(ethnic sex)
```

يحدد بيانات أزمنة العدّ. وفي هذا المثال، فإن المتغير *time* عبارة عن مقياس للوقت. والمتغير *nfail* يمثل عدد مرات الفشل التي حدثت عند الزمن *time*، ويمكننا إضافة متغيرات أخرى مثل المتغير *ncensor* (عدد المشاهدات التي تمت مراقبتها عند الزمن *time*) والمتغير *nenter* (عدد المدخلات عند الزمن *time*) وتجدر الإشارة إلى أن إضافة هذه المتغيرات اختيارية، كما أن المتغير *ethnic* والمتغير *sex* هما متغيران تصنيفيان آخران يُعرّفان المشاهدات في هذه البيانات.

```
.cttost
```

يقوم بتحويل بيانات أزمنة العد – والتي تم تحديدها مسبقاً باستخدام الأمر **ctset** – بحيث تكون في شكل أزمنة بقاء يمكن تحليلها باستخدام مجموعة أوامر **st***.

```
.sts graph
```

يقوم بإنشاء رسم بياني لدالة بقاء كابلان ميير Kaplan-Meier، ولمقارنة اثنتين أو أكثر من دوال البقاء – مثل مقارنة قيمة واحدة في كل متغير تصنيفي *sex* – نقوم باستخدام الخيار **by()**، وتكون تركيبة الأمر على شكل **sts graph, by(*sex*)** وللتعديل باستخدام انحدار كوكس على تأثيرات متغير مستقل مستمر مثل متغير العمر *age*، نقوم باستخدام الخيار **adjustfor()**، وتكون تركيبة الأمر على شكل **sts graph, by (*sex*) adjustfor(*age*)**. فالخيارات **by()** و**adjustfor()** تعمل بنفس الطريقة مع الأمر **sts list** والأمر **sts generate**.

```
.sts list
```

يقوم بإنشاء قائمة تحتوي على دالة بقاء كابلان ميير المقدّرة.

```
.sts test sex
```

يقوم باختبار التساوي بين دوال بقاء كابلان ميير في فئات المتغير *sex*.

```
.sts generate survfunc = S
```

يقوم بإنشاء متغير جديد تتم تسميته عشوائياً باسم *survfunc* يحتوي على دالة بقاء كابلان ميير المقدّرة.

```
.stcox x1 x2 x3
```

يقوم هذا الأمر بصياغة نموذج المخاطرة النسبي لكوكس، وانحدار الزمن إلى الفشل على المتغيرات التنبؤية الوهمية أو المستمرة *x3, x2, x1*.

```
.stcox x1 x2 x3, strata(x4) vce(robust)
.predict hazard, basechazard
```

يقوم هذا الأمر بصياغة نموذج المخاطرة النسبي لكوكس مقسّمة إلى طبقات باستخدام المتغير *x4*. الخيار **vce(robust)** يتطلّب تقدير الخطأ المعياري الموثوق، انظر الفصل (8) أو دليل المستخدم *User's Guide* للحصول على شرح كامل للأخطاء المعيارية الموثوقة. الأمر **predict** يقوم بحفظ دالة المخاطرة التراكمية الأساسية لمجموعة محددة، ويتم حفظها

كمتغير جديد باسم *hazard*. للحصول على خيارات أكثر قم بطباعة **help stcox postestimation**.

```
.stphplot, by(sex)
```

يقوم بإنشاء رسم بياني لـ ((ln(-ln(survival – مقابل ln(analysis time) لكل مستوى للمتغير النوعي *sex* من نموذج **stcox** السابق، تقريباً المنحنيات المتوازية تدعم فرضية نموذج كوكس التي تقول بأن نسبة المخاطرة لا تتغير مع الوقت. ولإجراء فحوصات أخرى لفرضيات كوكس يمكن القيام بها باستخدام الأمر **stcoxkm** (يُقارن منحنيات كوكس المتوقعة مع منحنيات البقاء المشاهدة لكابلان وميير) والأمر **estat phtest** (يقوم بالاختبار بناءً على بواقي شونفيلد Schoenfeld residuals). لمعرفة المزيد عن خيارات وتركيبة هذه الأوامر قم بطباعة **help stcox diagnostics**.

```
.streg x1 x2, dist(weibull)
```

يقوم بصياغة نموذج توزيع ويبل Weibull-distribution لانحدار نسبة الوقت إلى الفشل على متغير تنبؤي وهمي أو مستمر *x1* و*x2*.

```
.streg x1 x2 x3 x4, dist(exponential)
 vce(robust)
```

يقوم بصياغة نموذج التوزيع الأسّي لانحدار الوقت إلى الفشل على متغيرات تنبؤية وهمية أو مستمرة *x4, x3, x2, x1*، كما يقوم بحساب تقديرات الخطأ المعياري لاختلاف التباين الموثوق heteroskedasticity-robust، بالإضافة إلى توزيع ويبل، والتوزيع الأسّي والتوزيعات الأخرى للأمر **dist()** ومحددات الأمر **streg** بما فيها اللوغاريتم الطبيعي واللوغاريتم المنطقي وتوزيعات جاما المعيارية وتوزيعات جومبرتز Gompertz. وللحصول على مزيد من التفاصيل عن هذا الأمر وخياراته قم بطباعة الأمر **help streg**.

```
.stcurve, survival
```

بعد استخدام الأمر **streg**، فإن الأمر أعلاه يقوم بإنشاء رسم بياني لدالة البقاء لنتائج النموذج الذي تم حسابه بالأمر **streg** للقيم المتوسطة لكل متغيرات *x*.

```
.stcurve, cumhaz at(x3=50, x4=0)
```

بعد استخدام الأمر **streg**، فإن الأمر أعلاه يقوم بإنشاء رسم بياني لدالة المخاطرة التراكمية لنتائج النموذج الذي تم حسابه بالأمر **streg** للقيم المتوسطة للمتغيرات *x3, x2,x1* عند القيمة 50 أما المتغير *x4* فيكون عند القيمة 0.

```
.poisson count x1 x2 x3, irr exposure(x4)
```

يقوم بحساب انحدار بواسون لمتغير الأحداث المعدودة *count* (مفترضاً أنها تتبع توزيع بواسون) على المتغيرات المستقلة الوهمية أو المستمرة *x3, x2, x1* ، وسوف يتم عرض تأثيرات المتغيرات المستقلة كنسب لمعدل الوقوع (**irr**)، أما الخيار **exposure()** فهو يقوم بتحديد متغير يشير إلى كمية العرض في حالة عدم تساوي العرض لجميع المشاهدات؛ ويجب ملاحظة أن أي نموذج بواسون يفترض أن احتمالية الحدث تبقى ثابتة بغض النظر عن عدد مرات وقوع الحدث في كل مشاهدة. وإذا لم تبق الاحتمالية ثابتة، فإننا بدلاً من ذلك يجب أن نأخذ في الاعتبار استخدام الأمر **nbreg** (انحدار ذو حدين سالب) والأمر **gnbreg** (انحدار ذو حدين سالب معياري).

```
.glm count x1 x2 x3, link(log) family(poisson)
 exposure(x4) eform
```

يقوم بحساب الانحدار بنفس الطريقة المشار إليها في المثال أعلاه **poisson**، ولكن كنموذج خطي معياري (GLM)، الأمر **glm** يمكن أن يتناسب مع نموذج بواسون وذي الحدين السالب واللوغاريتمي والعديد من أنواع النماذج الأخرى، وهذا يعتمد على ما هي الخيارات **link()** (رابط الدالة) و**family()** (مجموعة التوزيع) التي تُستخدم.

بيانات أزمنة البقاء : Survival-Time Data

بيانات أزمنة البقاء تتضمن متغيرًا واحدًا على الأقل، وهذا المتغير يقوم بقياس كم يمضي من الوقت قبل وقوع حدث معين في كل مشاهدة. المراجع الإحصائية في العادة تقوم بتسمية هذا الحدث "فشل" بغض النظر عن المعنى

الحقيقي لهذه الكلمة، عند عدم وقوع الفشل في مشاهدة ما في بيانات أزمنة البقاء بنهاية عملية جمع هذه البيانات، فإن المشاهدة يجب أن يُقال عنها "مُراقبة". الأمر **stset** يحدد البيانات التي سوف يتم استخدامها في تحليل أزمنة البقاء، وذلك من خلال تحديد المتغير الذي يقوم بقياس الزمن (وإذا كان ضرورياً) تحديد المتغير الثنائي {1 ,0} كمؤشر لمعرفة ما إذا كانت المشاهدة "فشل" أو "مراقبة". البيانات يمكن أن تحتوي على أي رقم لمقياس آخر أو متغيرات نوعية، والأفراد (مثلاً المرضى في المستشفيات) يمكن تمثيلهم بواسطة أكثر من مشاهدة.

ولشرح استخدام الأمر **stset** سوف نبدأ بمثال من دراسة Selvin (1995:453) التي تحتوي على بيانات 51 شخصاً تم تشخيصهم على أنهم مصابون بمرض نقص المناعة HIV. البيانات توجد في الملف *aids.raw* وتظهر البيانات كما يلي:

```
 1   1   1  34
 2  17   1  42
 3  37   0  47
 (rows 4-50 omitted)
51  81   0  29
```

قيم العمود الأول (من الجهة اليسرى) تُظهر عدد الحالات (1، 2، 3، ... 51). العمود الثاني يوضح كم شهرًا مضى بعد التشخيص وقبل أن تظهر على الشخص أعراض مرض AIDS أو نهاية الدراسة (1، 17، 37، ...)، العمود الثالث يحتوي على 1 إذا كان الشخص ظهرت عليه أعراض مرض AIDS (الفشل) أو 0 إذا لم تظهر الأعراض في فترة نهاية الدراسة (مراقبة)، العمود الأخير يوضح أعمار الأشخاص عند وقت التشخيص.

يمكننا قراءة البيانات الخام في الذاكرة باستخدام الأمر **infile**، ثم نقوم بوصف المتغيرات والبيانات:

```
.infile case time aids age using
 C:\data\aids.raw, clear
```

```
.label variable case "Case ID number"
.label variable time "Months since HIV diagnosis"
.label variable aids "Developed AIDS symptoms"
.label variable age "Age in years"
.label data "AIDS (Selvin 1995:453)"
.compress
```

الخطوة التالية هي تحديد المتغير الذي يقوم بقياس الزمن، والذي يشير إلى الفشل أو المراقبة. وبالرغم من أنه ليس من الضروري مع هذا النوع من البيانات تحديد رقم مميز لكل حالة، فإننا سوف نقوم بذلك. الأمر **stset** يحدد المتغير الذي يقوم بقياس الزمن، وبالتالي سوف نحدد الفشل **failure()** كممثل وهمي يُحدد ما إذا كانت المشاهدة فشل (1) أو مراقبة (0)، وبعد استخدام الأمر **stset** سوف نقوم بحفظ البيانات في ملف بتنسيق ستاتا للحفاظ على هذه البيانات.

```
.stset time, failure(aids) id(case)

                id:  case
     failure event:  aids != 0 & aids < .
obs. time interval:  (time[_n-1], time]
 exit on or before:  failure

------------------------------------------------------------------------------
       51  total obs.
        0  exclusions
------------------------------------------------------------------------------
       51  obs. remaining, representing
       51  subjects
       25  failures in single failure-per-subject data
     3164  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =        97

.save aids.dta, replace
```

Stdescribe يعرض توصيفاً مختصرًا لكيفية تركيب بيانات أزمنة البقاء. لدينا في هذا المثال البسيط سجل واحد فقط لكل شخص، ولذلك فإن بعض هذه المعلومات غير ضرورية.

```
.stdescribe
```

```
         failure _d:  aids
   analysis time _t:  time
                 id:  case
```

Category	total	per subject mean	min	median	max
no. of subjects	51				
no. of records	51	1	1	1	1
(first) entry time		0	0	0	0
(final) exit time		62.03922	1	67	97
subjects with gap	0				
time on gap if gap	0	.	.	.	.
time at risk	3164	62.03922	1	67	97
failures	25	.4901961	0	0	1

الأمر **stsum** يقوم بحساب إحصائيات مختصرة، حيث يتضح أنه لـدينا 25 فشلاً من 3,164 شخص/ شهر، وهذا يعني أن معدل الحدوث هو 25 ÷ 3164 = 0.0079014 ، أما نسب زمن البقاء فسوف يتم اشتقاقها من دالـة البقاء لكابلان- مبير (الجزء التالي من هذا الفصل). الدالة تقوم بتقدير نحـو 25% فرصة للإصابة بمرض AIDS خلال فترة 41 شهرًا بعد التـشخيص، و50% خلال فترة 81 شهرًا، خلال الفترة التي تغطيها البيانات (نحـو 97 شهرًا) احتمالية الإصابة بمرض AIDS لم تصل لنسبة 75% ولـذا ليـست هناك نسبة ترتيبها 75.

.stsum

```
         failure _d:  aids
   analysis time _t:  time
                 id:  case
```

	time at risk	incidence rate	no. of subjects	Survival time 25%	50%	75%
total	3164	.0079014	51	41	81	.

إذا احتوت البيانات على متغير نوعي أو متغير تجميعي مثل الجـنس، فإنه يمكننا الحصول على إحصائيات مختصرة لزمن البقاء بـشكل منفـصل لكل مجموعة بواسطة الأمر التالي:

.stsum, by(sex)

الجزء التالي يشرح طرقاً أكثر منهجية للمقارنـــة بـــين أزمنـــة البقـــاء لمجموعتين أو أكثر.

بيانات حساب الزمن : Count-Time Data

بيانات زمن البقاء (st) بالملف *aids.dta* تحتوي علـــى معلومـــات عـــن حالات فردية (أشخاص أو أشياء) مع متغيرات تشير إلى زمن وقوع الفـــشل أو المراقبة لكل حالة فردية، هناك نوع آخر مختلف مـــن البيانـــات يـــسمى "حساب الزمن" تحتوي على بيانات إجمالية مع متغيرات تقوم بتعداد الحالات الفردية للفشل أو المراقبة عند الزمن *t*. فعلى سبيل المثال، ملـــف البيانـــات *diskdriv.dta* يحتوي على معلومات عن اختبار افتراضي لـــــ 25 قرصـــاً صلباً، كل الأقراص باستثناء 5 فقط فشلت قبل نهاية الاختبـــار عنـــد زمـــن 1,200 ساعة.

```
.use C:\data\diskdriv.dta, clear
.describe
```

```
Contains data from C:\data\diskdriv.dta
  obs:             6                          Count-time data on disk drives
 vars:             3                          1 Jul 2012 18:13
 size:            24
------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
hours           int    %8.0g                  Hours of continuous operation
failures        byte   %8.0g                  Number of failures observed
censored        byte   %9.0g                  Number still working
------------------------------------------------------------------------------
Sorted by:
```

```
.list
```

	hours	failures	censored
1.	200	2	0
2.	400	3	0
3.	600	4	0
4.	800	8	0
5.	1000	3	0
6.	1200	0	5

لتحديد بيانات على أنها بيانات حساب الـــزمن، يجب علينا تحديد متغير الوقت، ومتغير لعدد مرات الفشل، ومتغير لعدد مرات المراقبة على التوالي. فبعد استخدام الأمر **ctest** يقوم الأمر **cttost** تلقائياً بتحويل بيانـــات حـــساب الزمن إلى تنسيق زمن البقاء.

```
.ctset hours failures censored

 dataset name:  C:\data\diskdriv.dta
         time:  hours
     no. fail:  failures
     no. lost:  censored
    no. enter:  --                    (meaning all enter at time 0)


.cttost
     failure event:  failures != 0 & failures < .
obs. time interval:  (0, hours]
 exit on or before:  failure
            weight:  [fweight=w]

------------------------------------------------------------------------------
        6  total obs.
        0  exclusions
------------------------------------------------------------------------------
        6  physical obs. remaining, equal to
       25  weighted obs., representing
       20  failures in single record/single failure data
    19400  total analysis time at risk, at risk from t =         0
                                 earliest observed entry t =         0
                                      last observed exit t =      1200


.list

     +------------------------------------------------------------+
     | hours   failures   censored   w   _st   _d     _t   _t0 |
     |------------------------------------------------------------|
  1. |   200          1          0   2     1    1    200     0 |
  2. |   400          1          0   3     1    1    400     0 |
  3. |   600          1          0   4     1    1    600     0 |
  4. |   800          1          0   8     1    1    800     0 |
  5. |  1000          1          0   3     1    1   1000     0 |
     |------------------------------------------------------------|
  6. |  1200          0          5   5     1    0   1200     0 |
     +------------------------------------------------------------+


.stdescribe
```

```
        failure _d:  failures
  analysis time _t:  hours
            weight:  [fweight=w]
```

Category	unweighted total	unweighted mean	min (per subject)	unweighted median	max
no. of subjects	6				
no. of records	6	1	1	1	1
(first) entry time		0	0	0	0
(final) exit time		700	200	700	1200
subjects with gap	0				
time on gap if gap	0				
time at risk	4200	700	200	700	1200
failures	5	.8333333	0	1	1

الأمر **cttost** يقوم بتحديد مجموعة من الأوزان التكرارية *w* في البيانات المستخرجة ذات تنسيق -st، كما أن الأوامر التي تبدأ بـ **st*** تقوم تلقائياً بالتعرّف على هذه الأوزان واستخدامها في أي تحليل لزمن البقاء، ولذلك فإن البيانات التي رأيناها سابقاً تحتوي على 25 مشاهدة (25 قرصاً صلباً) بدلاً من الوضع السابق وهو 6 (ست فترات زمنية).

.stsum

```
        failure _d:  failures
  analysis time _t:  hours
            weight:  [fweight=w]
```

	time at risk	incidence rate	no. of subjects	Survival time 25%	50%	75%
total	19400	.0010309	25	600	800	1000

دوال بقاء كابلان ـ مير : Kaplan-Meier Survivor Functions

بافتراض أن n تمثل عدد المشاهدات التي لا يوجد بها فشل أو مراقبة عند بداية الفترة الزمنية t، و d_t تمثل عدد مرات الفشل التي حدثت لتلك

المشاهدات خلال الفترة الزمنية t، مُقدِّر كابلان – ميير للبقاء بعد الفترة t هو ناتج احتمالات البقاء عند الزمن t والفترات السابقة:

$$S(t) = \prod_{j=t0}^{t} \{(n_j - d_j) \div n_j\} \qquad [10.1]$$

فمثلاً في بيانات مرض نقص المناعة AIDS التي رأيناها سابقاً، فإن شخصاً واحدًا من 51 شخصاً ظهرت عليه الأعراض بعد شهر واحد فقط من التشخيص، ولم تكن هناك مشاهدات مراقبة عند هذه الفترة المبكرة، ولذلك فإن احتمالية "البقاء" (أي عدم ظهور أعراض AIDS) بعد الفترة $time = 1$ هو

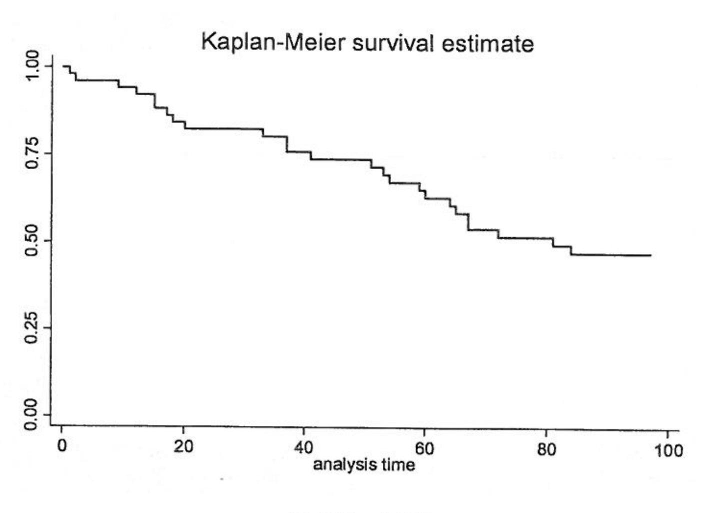
$$S(1) = (51 - 1) / 51 = 0.9804$$

مريض ثان ظهرت عليه الأعراض عند الفترة $time = 2$، ومريض ثالث عند الفترة $time = 9$:

$$S(2) = 0.9804 \times (51 - 1) / 51 = 0.9608$$
$$S(9) = 0.9608 \times (51 - 1) / 51 = 0.9412$$

وعند إنشاء رسم بياني للمرضى $S(t)$ مع الزمن t ، فإن منحنى البقاء لكابلان – ميير يظهر كما في الشكل (1.10). برنامج ستاتا يقومٍ برسم مثل هذه الأشكال البيانية بشكل تلقائي باستخدام الأمر **sts graph** فمثلاً:

```
.use C:\data\aids, clear
.sts graph
```



الشكل (1.10)

المثال الثاني الخاص بدوال البقاء – سوف نتحول إلى البيانات الموجـودة بالملف *smoking1.dta* – والتي تم الحصول عليها مـن دراسـة Rosner (1995)، المشاهدات الموجودة بالبيانات عبارة عن 234 مدخناً سابقاً يحاولون الإقلاع عن التدخين، أغلب المدخنين لم ينجحوا في الإقلاع عن التدخين، المتغير *days* يقـوم بتسجيل عدد الأيام ما بين الإقلاع عن التدخين والعودة إليه من جديـد. وغطـت الدراسة مدة سنة واحدة، المتغير *smoking* يشير إلى ما إذا كان الشخص قد عـاد إلى التدخين قبل نهاية مدة الدراسة ("فشل"، *smoking* = 1) أو لم يعـد للتـدخين ("مراقبة"، *smoking* = 0)، مع البيانات الجديدة يُفترض أن نبدأ باسـتخدام الأمـر **stset** لجعل البيانات جاهزة لتحليل زمن البقاء.

```
.use C:\data\smoking1.dta, clear
.describe
Contains data from C:\data\smoking1.dta
  obs:           234                          Smoking (Rosner 1995:607)
 vars:             8                          2 Jul 2012 06:11
 size:         2,808
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
id              int    %9.0g                  Case ID number
days            int    %9.0g                  Days abstinent
smoking         byte   %9.0g                  Resumed smoking
age             byte   %9.0g                  Age in years
sex             byte   %9.0g       sex        Sex (female)
cigs            byte   %9.0g                  Cigarettes per day
co              int    %9.0g                  Carbon monoxide x 10
minutes         int    %9.0g                  Minutes elapsed since last cig
-------------------------------------------------------------------------------
Sorted by:

.stset days, failure(smoking)
     failure event:  smoking != 0 & smoking < .
obs. time interval:  (0, days]
 exit on or before:  failure

-------------------------------------------------------------------------------
      234  total obs.
        0  exclusions
-------------------------------------------------------------------------------
      234  obs. remaining, representing
      201  failures in single record/single failure data
    18946  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =       366
```

الدراسة تتضمن 110 رجال و124 امرأة، معدلات الوقوع لكلا الجنسين يبدو أنها متشابهة.

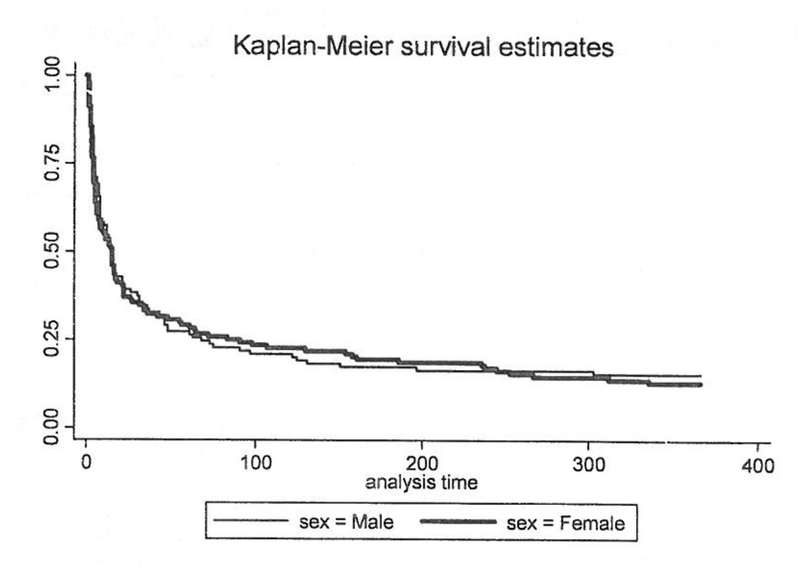
```
.stsum, by(sex)

        failure _d:  smoking
  analysis time _t:  days
```

sex	time at risk	incidence rate	no. of subjects	Survival time 25%	50%	75%
Male	8813	.0105526	110	4	15	68
Female	10133	.0106582	124	4	15	83
total	18946	.0106091	234	4	15	73

الشكل (2.10) يؤكد هذا التشابه، حيث يُظهِر الشكل اختلافاً بسيطاً بين دوال البقاء للرجال والنساء، حيث عاد كلا الجنسين للتدخين من جديد في نفس الوقت تقريباً. احتمالات البقاء للأشخاص غير المدخنين تنخفض بشكل كبير خلال فترة الـ 30 يوماً الأولى بعد الإقلاع عنه، لكلا الجنسين هناك احتمال أقل من 15% للبقاء كشخص غير مدخن بعد انقضاء سنة.

```
.sts graph, by(sex) plot1opt(lwidth(medium))
 plot2opt(lwidth(thick))
```



الشكل (2.10)

يمكننا اختبار التساوي في دوال البقاء باستخدام اختبار لوغاريتم الرتب، وليس غريباً أن هذا الاختبار لم يجد أي اختلاف ذا معنوية إحصائية ($p = 0.6772$) للعودة للتدخين بين الرجال والنساء.

```
.sts test sex

              failure _d:  smoking
        analysis time _t:  days


        Log-rank test for equality of survivor functions

       sex    |  Events         Events
              |  observed       expected
       -------+---------------------------
       Male   |       93           95.88
       Female |      108          105.12
       -------+---------------------------
       Total  |      201          201.00

                    chi2(1) =       0.17
                    Pr>chi2 =     0.6772
```

نماذج المخاطر النسبية لكوكس : Cox Proportional Hazard Models

نماذج الانحدار تسمح لنا بأن نوسع تحليل البقاء، ونختبر تأثيرات المتغيرات التنبؤية الطبقية أو المستمرة المتعددة. إحدى الطرق الأكثر استخداماً للقيام بذلك تُعرف باسم انحدار كوكس، وهذه الطريقة تستخدم نموذج مخاطرة نسبياً. ومعدل المخاطرة للفشل عند الزمن t يتم تعريفها بأنها معدل الفشل عند الزمن t بين الذين بقوا حتى الزمن t :

$$h(t) = \frac{\text{احتمال الفشل بين الوقت } t \text{ والوقت } t + \Delta t}{(\text{احتمال الفشل بعد الوقت } t)\ (\Delta t)} \qquad [2.10]$$

نقوم بصياغة نموذج لمعدل المخاطرة كدالة لخط الخطر الأساسي (h_0) عند الوقت t وتأثيرات واحد أو أكثر من متغيرات x :

$$h(t) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k) \qquad [a10.3]$$

وهذا يكافئ

$$\ln[h(t)] = \ln[h_0(t)]\exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k) \qquad [b10.3]$$

"**خط الخطر الأساسي**" يعني أن الخطر لمشاهدة ما مع كل متغيــرات x يساوي صفرًا. انحدار كوكس يقوم بتقدير هذا الخــط بطريقــة لامعلميــة، ويحصل على تقدير الأرجحية العظمى لــ β لمعلمات المعادلة [10.3]. الأمر **stcox** هو إجراء يقوم بعرض نسب المخاطرة والتي هي عبارة عن تقديرات exp(β)، وهي تشير إلى التغيرات النسبية إلى معدل خط الخطر الأساسي.

هل يؤثر العمر على بداية أعراض مرض نقص المناعة AIDS؟ ملــف البيانات *aids.dta* يتضمن معلومات تناقش هذا السؤال. يجــب ملاحظــة أن الأمر **stcox** – يختلف عن أغلب أوامر صياغة نماذج ستاتا – حيــث إننــا سنقوم بإدراج المتغير أو المتغيرات المستقلة فقط، أما المتغيــرات التابعــة، ومتغيرات الوقت، ومتغيرات المراقبة فإن ستاتا يفهمها بشكل تلقــائي عنــد استخدام الأمر **stset**.

```
.stcox age, nolog

         failure _d:  aids
   analysis time _t:  time
                 id:  case

Cox regression -- Breslow method for ties

No. of subjects =          51                  Number of obs   =        51
No. of failures =          25
Time at risk    =        3164
                                               LR chi2(1)      =      5.00
Log likelihood  =  -86.576295                  Prob > chi2     =    0.0254
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf. Interval]	
age	1.084557	.0378623	2.33	0.020	1.01283	1.161363

قد نقوم بتفسير معدل المخاطرة المقدّر وهو 1.084557 مع الإشارة إلى شخصين مصابين بمرض HIV أعمارهما a وa+1، الشخص الأكبر سناً يكون أكثر عرضة لظهور أعراض مرض AIDS بنسبة 8.5% خلال فترة أقــصر (المعدل الخاص بالمخاطر هو 1.084557)، هذا المعدل يختلف اختلافــاً ذا معنوية إحصائية (p = 0.02) عن 1. إذا كنا نريد شرح نتائجنا لفرق خمــس سنوات في السن، فإننا سوف نتساءل عن معدل المخاطر للأس 5 كما يلي:

```
.display exp(_b[age])^5
 1.5005865
```

ولذلك، فإن خطر بداية الإصابة بمرض AIDS أعلى بنحو 50% عندما يكون الشخص الثاني أكبر بخمس سنوات من الشخص الأول، وبدلاً من ذلك فإنه بالإمكان أن نجد نفس الشيء (ونحصل على فترة ثقة جديدة) بتكرار الانحدار بعد تحوير متغير العمر *age*، بحيث يتم القياس على أساس 5 سنوات، الخيار أدناه **nolog noshow** يمنع عرض السجل المتكرر وشرح البيانات -st

```
.generate age5 = age/5
.label variable age5 "age in 5-year units"
.stcox age5, nolog noshow
```

```
Cox regression -- Breslow method for ties

No. of subjects =          51                 Number of obs    =        51
No. of failures =          25
Time at risk    =        3164
                                              LR chi2(1)       =      5.00
Log likelihood  =  -86.576295                 Prob > chi2      =    0.0254
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age5	1.500587	.2619305	2.33	0.020	1.065815	2.112711

مشابهاً للانحدار العادي، فإن نماذج كوكس يمكن أن يكون لها أكثر من متغير مستقل واحد. وباستخدام ملف البيانات *heart.dta* الذي يحتوي على بيانات زمن بقاء من دراسة Selvin (1995) لـ 35 مريضاً لديهم مستويات مرتفعة جداً من الكليسترول، فإن متغير الزمن *time* يوضح عدد الأيام لكل مريض كان تحت المتابعة. والمتغير *coronary* يشير إلى ما إذا كانت النوبة القلبية قد حدثت في نهاية فترة المتابعة (*coronary* = 1) أو لم تحدث (*coronary* = 0). البيانات تتضمن أيضاً مستويات الكليسترول وعوامل أخرى يُعتقد بأن لها تأثيرًا على مرض القلب، ملف البيانات *heart.dta* تم إعداده لتحليل زمن البقاء عن طرق الأمر **stset *time*, failure(*coronary*)**، ولذلك يمكننا أن نقوم باستخدام تحليل **st** مباشرة.

```
.describe patient-ab

              storage  display    value
variable name   type   format     label      variable label
---------------------------------------------------------------------------
patient         byte   %9.0g                 Patient ID number
time            int    %9.0g                 Time in days
coronary        byte   %9.0g                 Coronary event (1) or none (0)
weight          int    %9.0g                 Weight in pounds
sbp             int    %9.0g                 Systolic blood pressure
chol            int    %9.0g                 Cholesterol level
cigs            byte   %9.0g                 Cigarettes smoked per day
ab              byte   %9.0g                 Type A (1) or B (0) personality
```

```
.stdescribe

        failure _d:  coronary
  analysis time _t:  time

                                   |-------------- per subject --------------|
Category                   total        mean         min     median        max
--------------------------------------------------------------------------------
no. of subjects               35
no. of records                35           1           1          1          1

(first) entry time                         0           0          0          0
(final) exit time                   2580.629         773       2875       3141

subjects with gap              0
time on gap if gap             0
time at risk               90322    2580.629         773       2875       3141

failures                       8    .2285714           0          0          1
--------------------------------------------------------------------------------
```

انحدار كوكس وجد أن مستوى الكليسترول والتدخين معاً يزيدان بشكل كبير من مخاطر النوبة القلبية. وبعكس المتوقع، فإن زيادة الوزن تؤدي إلى تخفيض مخاطر النوبة القلبية، أما ضغط الدم ونوع الشخصية A/B ليس لهما تأثير ذو معنوية إحصائية.

```
.stcox weight sbp chol cigs ab, noshow nolog
```

```
Cox regression -- no ties

No. of subjects =           35                  Number of obs   =          35
No. of failures =            8
Time at risk    =        90322
                                                LR chi2(5)      =       13.97
Log likelihood  =   -17.263231                  Prob > chi2     =      0.0158
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
weight	.9349336	.0305184	-2.06	0.039	.8769919	.9967034
sbp	1.012947	.0338061	0.39	0.700	.9488087	1.081421
chol	1.032142	.0139984	2.33	0.020	1.005067	1.059947
cigs	1.203335	.1071031	2.08	0.038	1.010707	1.432676
ab	3.04969	2.985616	1.14	0.255	.4476492	20.77655

بعد تقدير النموذج يمكننا توقع **predict** متغيرات جديدة تحتفظ بخط الخطر الأساسي التراكمي ودوال البقاء. وبما أن الخط الأساسي يشير إلى الوضعية التي تكون عندها كل متغيرات *x* تساوي صفرًا، ويجب علينا أولاً القيام بتمركز بعض المتغيرات حتى تكون قيم صفر لها معنى، فالمريض الذي وزنه صفر باوند أو ضغط دمه صفر ليس له فائدة عند إجراء عملية المقارنة، وباستخدام أصغر القيم في البيانات الموجودة لدينا، فإنه يمكننا أن نقوم بعملية تحوير لمتغير الوزن *weight* بحيث يشير الصفر إلى 120 باوند، وصفر للمتغير *sbp* يشير إلى 105، وصفر للمتغير *chol* يشير إلى 340:

```
.summarize patient- ab
```

Variable	Obs	Mean	Std. Dev.	Min	Max
patient	35	18	10.24695	1	35
time	35	2580.629	616.0796	773	3141
coronary	35	.2285714	.426043	0	1
weight	35	170.0857	23.55516	120	225
sbp	35	129.7143	14.28403	104	154
chol	35	369.2857	51.32284	343	645
cigs	35	17.14286	13.07702	0	40
ab	35	.5142857	.5070926	0	1

```
.replace weight = weight - 120
.replace sbp = sbp - 105
.replace chol = chol - 340
```

```
.summarize patient - ab
```

Variable	Obs	Mean	Std. Dev.	Min	Max
patient	35	18	10.24695	1	35
time	35	2580.629	616.0796	773	3141
coronary	35	.2285714	.426043	0	1
weight	35	50.08571	23.55516	0	105
sbp	35	24.71429	14.28403	-1	49
chol	35	29.28571	51.32284	3	305
cigs	35	17.14286	13.07702	0	40
ab	35	.5142857	.5070926	0	1

قيم صفر لكل متغيرات x يكون لها معنى حقيقي الآن. ولإنشاء متغيرات جديدة تحتوي على الخط الأساسي للبقاء، وتقديرات دالة الخطر التراكمـي، فإننا نقوم بتكرار الانحدار، ونتبع ذلك بأمرين **predict** كما يلي:

```
.stcox weight sbp chol cigs ab, noshow nolog
.predict hazard, basechazard
.predict survivor, basesurv
Cox regression -- no ties

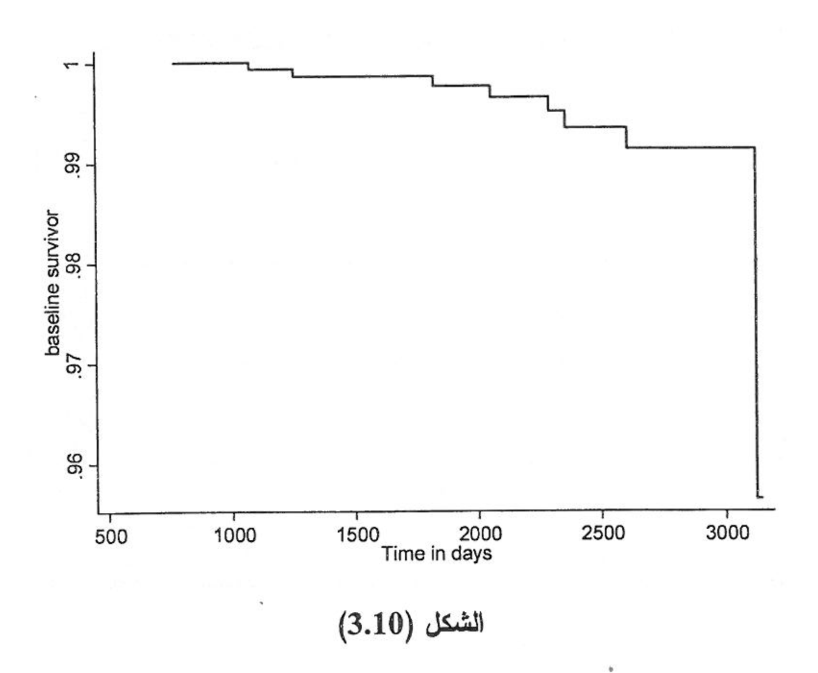
No. of subjects =          35                Number of obs   =        35
No. of failures =           8
Time at risk    =       90322
                                             LR chi2(5)      =     13.97
Log likelihood  =  -17.263231                Prob > chi2     =    0.0158
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
weight	.9349336	.0305184	-2.06	0.039	.8769919	.9967034
sbp	1.012947	.0338061	0.39	0.700	.9488087	1.081421
chol	1.032142	.0139984	2.33	0.020	1.005067	1.059947
cigs	1.203335	.1071031	2.08	0.038	1.010707	1.432676
ab	3.04969	2.985616	1.14	0.255	.4476492	20.77655

ويجب ملاحظة أن إنشاء ثلاثة متغيرات x ليس له تأثير علـى نسب المخاطرة والأخطاء المعيارية وغيرها من الإحصائيات الأخـرى. أوامـر **predict** تقوم بإنشاء متغيرات جديدة ويتم تسميتها عشوائياً بأسـماء *hazard,* *survivor*، ولإنشاء رسم بياني يمثل دالة الخط الأساسي للبقاء، فإننـا نقـوم

بإنشاء رسم بياني للمتغير *survivor* مع المتغير *time*، ثم نقوم بتوصيل النقاط بطريقة تشبه السلالم كما تظهر في الشكل (3.10).

```
.graph twoway line survivor time,
 connect(stairstep) sort
```



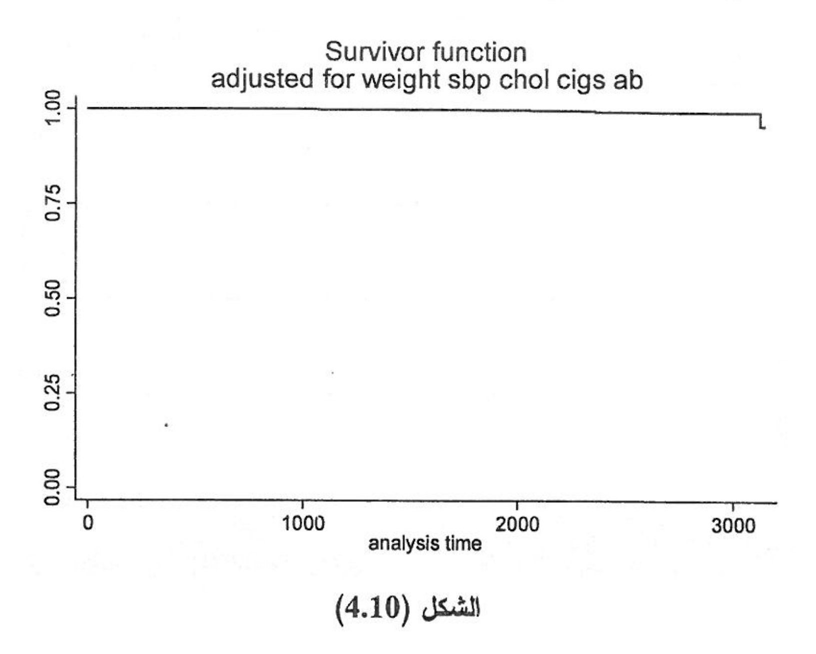
الشكل (3.10)

دالة الخط الأساسي للبقاء – والتي تصف احتمالات البقاء للمرضى الذين وزنهم صفر (120 باوند)، وضغط الدم صفر (105) والكليسترول صفر (340)، ولايدخنون، وهم أشخاص من نوع الشخصية B– تنخفض مع الوقت، وبالرغم من أن هذا الانخفاض يبدو كبيرًا جداً في الجانب الأيمن من الرسم، فإنه من الملاحظ أن الانخفاض حقيقةً كان من 1 إلى نحو 0.96. وبالنظر إلى القيم الأقل تفضيلاً للمتغيرات التنبؤية، فإن احتمالات البقاء كان من المفترض أن تكون أكثر انخفاضاً.

نفس الشكل البياني لدالة الخط الأساسي للبقاء كان يمكن الحصول عليها بطريقة أخرى بدون الأمر **stcox**، وهذه الطريقة تم استخدامها لإنشاء الشكل

(4.10) حيث إنها تستخدم الأمر **sts graph** مع الخيارات **adjustfor()**، ثم بعد ذلك يتم إدراج المتغيرات التنبؤية.

```
.sts graph, adjustfor(weight sbp chol cigs ab)
```

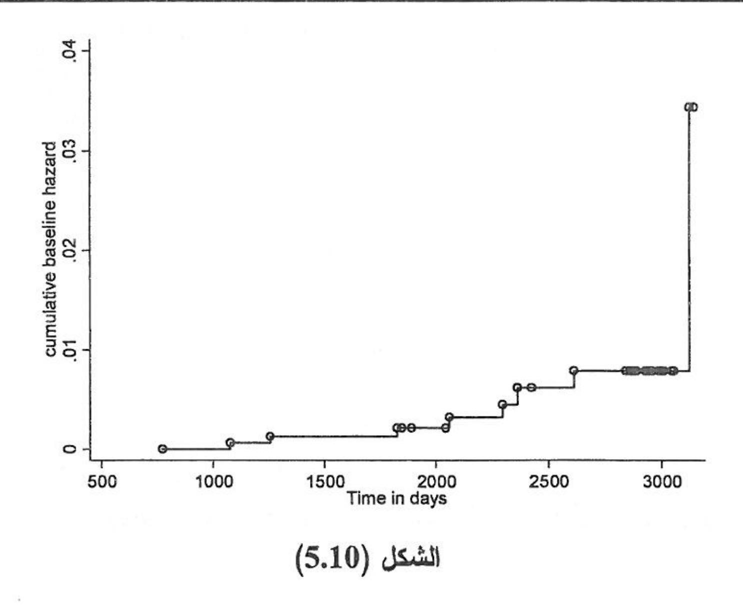


الشكل (4.10)

الشكل (4.10)– يختلف عن الشكل (3.10) – حيث إنه يعرض دالة البقاء العادية مع تقسيم قياس المحور العمودي ليكون من 0 إلى 1، مع وجود هذا الاختلاف في القياس بالمحور العمودي. الشكل (3.10) والشكل (4.10) يعرضان نفس المنحنى.

الشكل (5.10) يعرض رسماً بيانياً لخط الخطر الأساسي التراكمي مع الوقت باستخدام المتغير *hazard* والذي تم إنشاؤه بالأمر **stcox**، الشكل يوضح بأن خط الخطر الأساسي التراكمي قد ازداد عند 8 نقاط (لأن 8 مرضى "فشلوا" أو حدثت لهم نوبة قلبية) من 0 تقريباً إلى 0.033.

```
.graph twoway connected hazard time,
 connect(stairstep) sort msymbol(Oh)
```



الشكل (5.10)

انحدار ويبل Weibull والانحدار الأُسّي :

Exponential and Weibull Regression

انحدار كوكس يقوم بتقدير دالة خط الخطر الأساسي بدون الإشارة إلى أي توزيع نظري. هناك العديد من المداخل المعلمية الأخرى تبدأ من افتراضات بأن أزمنة البقاء تتبع توزيعات نظرية معروفة. هذه التوزيعات تتضمن التوزيع الأسّي، وتوزيع ويبل، والتوزيع اللوغاريتمي الطبيعي، والتوزيع اللوغاريتمي الثنائي، وتوزيع جومبرتز، أو توزيع جاما المعياري. النماذج التي يتم إنشاؤها بناءً على أي من هذه التوزيعات يمكن استخدامها مع الأمر **streg**، ومثل هذه النماذج لها نفس الصيغة العامة لانحدار كوكس (المعادلتان [2.10] و [3.10]) ولكنها تُعرِّف الخط الأساسي للخطر $h_0 = (t)$ بطريقة مختلفة. هناك مثالان عن ذلك سوف يتم تناولهما في هذا الجزء.

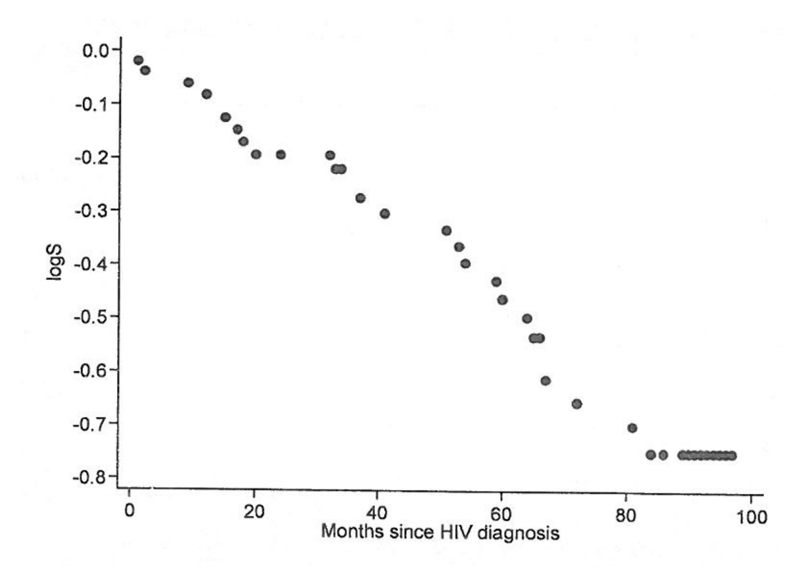
إذا حدث الفشل بطريقة مستقلة مع خطر ثابت عند أزمنة بقاء، وهذا الحدث يتبع توزيع أسّي، فإنه بالإمكان تحليله باستخدام الانحدار الأسّي. الخطر الثابت يعني الأفراد المشاركين في الدراسة ليسوا أكثر احتمالاً أو أقل عرضة للفشل في آخر مدة المشاهدة عنهم في بدايتها. لفترة طويلة هذه الفرضية لم تكن مُبررة للآلات والكائنات الحية، ولكن يمكن إيقاف تأثيرها إذا كانت فترة

المشاهدة تغطي جزءًا بسيطاً نسبياً من فترات الحياة، النموذج الأسّي يعتمد على أن لوغاريتمات دالة البقاء – $\ln(S)(t))$ – ترتبط خطيّاً مع t.

المدخل المعلمي الثاني الشائع وهو انحدار ويبل، والذي يعتمد على توزيع أكثر عموماً، وهو توزيع ويبل، هذا لايتطلب أن تكون معدلات الفشل ثابتة، ولكن يسمح لهذه المعدلات بالزيادة أو النقص بشكل سلس خلال فترة من الزمن، نموذج ويبل يفترض بأن $\ln(-\ln(S(t)))$ هو دالة خطية لـ $\ln(t)$.

الرسومات البيانية تعتبر وسيلة تشخيصية مفيدة لمعرفة مدى ملاءمة نماذج ويبل أو النماذج الأسّية. فمثلاً وبالعودة إلى ملف بيانات *aids.dta* نقوم بإنشاء رسم بياني (الشكل 6.10) لـ $\ln(S(t))$ مقابل الزمن وذلك بعد إنشاء تقديرات كابلان-مبير لدالة البقاء $S(t)$، توصيفات المحور العمودي في الشكل (6.10) تم إعطاؤها رقمين ثابتين مع تنسيق يعرض رقماً واحدًا بعد الفاصلة (%2.1f) ويكون اتجاه هذه الأرقام أفقياً، وذلك حتى يمكن قراءتها بوضوح.

```
.sts gen S = S
.generate logS = ln(S)
.graph twoway scatter logS time,
 ylabel(-.8(.1)0, format(%2.1f) angle (horizontal))
```



الشكل (6.10)

نمط الشكل (6.10) يبدو خطّياً بطريقة ما مما يدفعنا إلى محاولة الانتقال إلى انحدار أسّي:

```
.streg age, dist(exponential) nolog noshow

Exponential regression -- log relative-hazard form

No. of subjects =           51                  Number of obs   =        51
No. of failures =           25
Time at risk    =         3164
                                                LR chi2(1)      =      4.34
Log likelihood  =   -59.996976                  Prob > chi2     =    0.0372
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf. Interval]	
age	1.074414	.0349626	2.21	0.027	1.008028	1.145172
_cons	.0006811	.0007954	-6.24	0.000	.000069	.0067191

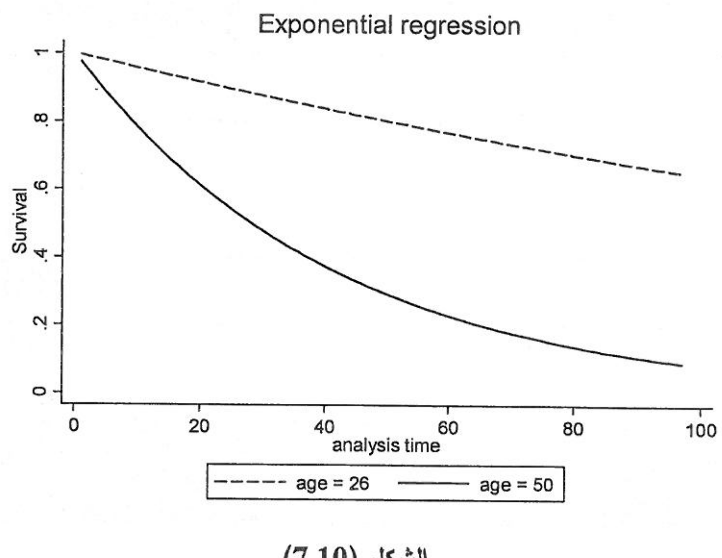
معدل الخطر (1.074) والخطأ المعياري (0.035) تم تقديرهما بواسطة الانحدار الأسّي أعلاه لا تختلفان بشكل كبير عن نظائرها في انحدار كوكس الذي قمنا به سابقاً (1.085 و 0.038). هذا التشابه يعكس درجة الترابط بين دالة الخطر التجريبية والخطر الثابت والذي تم إنشاؤه بواسطة التوزيع الأسّي، وبناءً على هذا النموذج الأسّي، فإن خطر إصابة الشخص بمرض HIV وظهور أعراض مرض AIDS تزداد بنحو 7.4% مع كل سنة زيادة في العمر.

بعد إجراء الحسابات بالأمر **streg** سوف نقوم بإنشاء رسم بياني باستخدام الأمر **stcurve** لنموذج الخطر التراكمي ودالة الخطر أو دالة البقاء، والوضع الافتراضي هو أن يقوم الأمر **stcurve** برسم هذه المنحنيات مع الاحتفاظ بكل قيم المتغيرات *x* في النموذج عند متوسطاتها. ويمكننا تحديد قيم *x* الأخرى باستخدام الخيار **at()**، وتجب الملاحظة أن بيانات الأشخاص الموجودة في الملف *aids.dta* تتراوح ما بين 26 إلى 50 سنة، ويمكننا إنشاء رسم بياني لدالة البقاء عندما يكون العمر *age* = 26 وذلك باستخدام أمر مثل:

```
.stcurve, surviv at(age = 26)
```

الشكل البياني يمكن أن يصبح أكثر تنسيقاً باستخدام الخيار ()at1 والخيار ()at2 لعرض منحنى البقاء باستخدام نوعين من قيم *x* مثل أعلى وأقل قيمة لمتغير *age* كما يلي:

```
.stcurve, survival at1(age = 26) at2(age = 50)
 lpattern(dash solid)
```



الشكل (7.10)

الشكل (7.10) يعرض منحنى البقاء المتوقع (للتحول من HIV ويتم تشخيصه AIDS) وهو ينخفض بدرجة كبيرة بين المرضى المُسنّين، حيث معدل المخاطرة الجوهري بالنسبة للعمر *age* وهو أكبر من 1 في جدول التوزيع الأسّي له نفس المعنى، ولكن باستخدام الأمر **stcurve** مع الخيارين ()**at1** و()**at2** فإن القيم تعطي شرحاً مرئياً للتأثير أكثر قوة من الجدول. هذه الخيارات تعمل بطريقة مشابهة مع الأنواع الثلاثة لأمر الرسم البياني **stcurve**

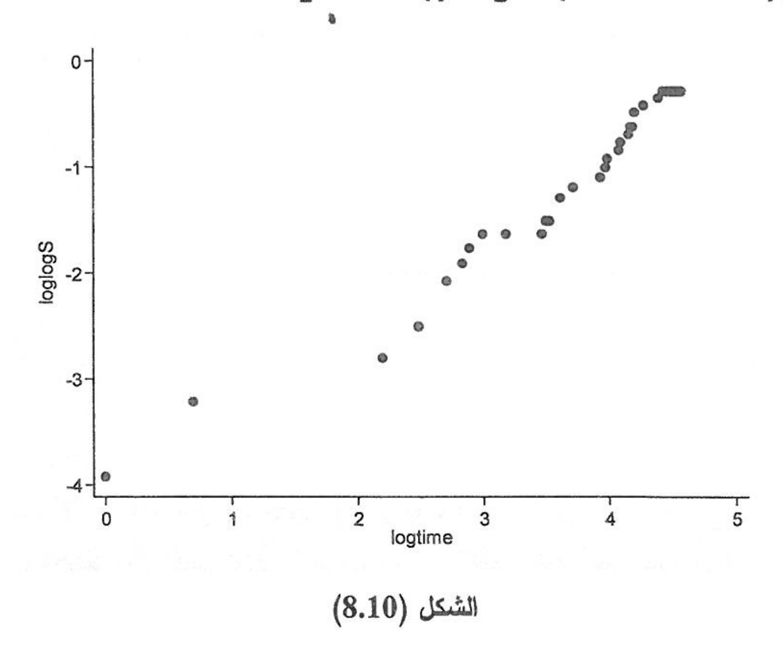
stcurve, survival دالة البقاء

stcurve, hazard دالة الخطر

stcurve, cumhaz دالة الخطر التراكمي

وبدلاً من التوزيع الأسّي يمكن للأمر **streg** صياغة نماذج بقاء تعتمد على توزيع ويبل. وتوزيع ويبل قد يبدو غير خطّي في الرسم البياني لـ $\ln(S(t))$ مقبل الزمن t، ولكن يُفترض أن يظهر بشكل خطّي في الرسم البياني لـ $\ln(-\ln(S(t)))$ مقابل $\ln(t)$ مشابهاً للشكل (8.10). ومن ناحية أخرى، فإن التوزيع الأسّي سوف يظهرخطّياً في كلا الشكلين البيانيين مع ميل يساوي 1 في $\ln(-\ln(S(t)))$ مقابل الرسم البياني لـ $\ln(t)$. وفي الحقيقة فإن بيانات المرضى في الشكل (8.10) قريبة من العلاقة الخطيّة مع ميل يساوي 1، وهذا يعني أن النموذج الأسّي السابق كان كافياً.

```
.generate loglogS = ln(-ln(S))
.generate logtime = ln(time)
.graph twoway scatter loglogS logtime,
                  ylabel(,angle(horizontal))
```



الشكل (8.10)

ولذلك، فإننا لانحتاج إلى تعقيدات أكثر لأي نموذج من نماذج ويبل مع هذه البيانات والنتائج تم عرضها بالجدول أدناه.

```
.streg age, dist(weibull) noshow nolog
```

```
Weibull regression -- log relative-hazard form

No. of subjects =           51                  Number of obs   =         51
No. of failures =           25
Time at risk    =         3164
                                                LR chi2(1)      =       4.68
Log likelihood  =   -59.778257                  Prob > chi2     =     0.0306
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	1.079477	.0363509	2.27	0.023	1.010531	1.153127
_cons	.0003313	.0005415	-4.90	0.000	.0000135	.0081564
/ln_p	.1232638	.1820858	0.68	0.498	-.2336179	.4801454
p	1.131183	.2059723			.7916643	1.616309
1/p	.8840305	.1609694			.6186934	1.263162

انحدار ويبل يحسب معدل المخاطرة ويقدرها بـــ (1.079)، وهو يتوسط بين نتائج كوكس السابقة والنتائج الأسّية. الاختلاف الأكثر وضوحاً في النماذج السابقة، هو وجود ثلاثة خطوط جديدة في أسفل الجدول، وهي تشير إلى معلمة شكل توزيع ويبل p. وعندما تكون قيمة p تساوي 1، فإن ذلك يعني أنها تتعلق بنموذج أسّي، والمخاطر لاتتغير بمرور الزمن. أما إذا كانت $p>1$ فهذا يشير إلى أن المخاطر تزداد مع الوقت، وفي حالة أن $p<1$ فهذا يعني أن المخاطر تنخفض، وفترة الثقة 95% لمعلمة توزيع ويبل p تتراوح ما بين 0.79 و1.62 ولذلك فليس لدينا أي سبب لرفض النموذج الأسّي ($p = 1$) هنا، نموذج ويبـــل يركز على $\ln(p), p, 1/p$ هذا التركيز يتم بشكل مختلف، ولكن رياضياً متساوي، ولذا فإن برنامج ستاتا يوفر هذه الطرق الثلاث، فالأمر **stcurve** يقوم بإنـــشاء رسم بياني لدوال المخاطرة التراكمية، أو دوال المخاطر، أو دوال البقاء بعـــد الأمـــر **streg, dist(weibull)** كالـــذي تـــم القيـــام بـــه بعـــد الأمـــر **streg, dist(exponenetial)** أو النماذج الأخرى لـــ **streg**.

الانحدار الأسّي أو انحدار ويبل أكثر تفضيلاً من انحدار كوكس عنـــدما تكون أزمنة البقاء تتبع توزيعاً أسّياً أو توزيع ويبل. ولكن عند صياغة نماذج الانحدار هذه بطريقة خاطئة، فإننا سوف نحصل على نتائج مضللة. انحـــدار كوكس– والذي لايعتمد على أي افتراضات سابقة حول شكل التوزيع – يبقى الأداة المفضلة في حالات كثيرة.

بالإضافة إلى النماذج الأسّية ونماذج ويبل، فإنه بالإمكان استخدام الأمر **streg** في صياغة العديد من النماذج التي تعتمد علـــى توزيـــع جـــومبرتز، والتوزيع اللوغاريتمي الطبيعي، والتوزيع اللوغاريتمي الثنائي، وتوزيع جاما المعياري. وللحصول على معلومات أكثر حول تركيبة الأمر وخياراته، قــم بطباعة الأمر **help streg** أو الاطلاع على دليل المستخدم *Survival Analysis and Epidemiological Tables Reference Manual*.

انحدار بواسون : Poisson Regression

إذا كانت الأحداث تقع بشكل مستقل وبمعدل ثابت، فإن عـــدد الأحـــداث خلال فترة معينة من الزمن، تتبع توزيع بواسون. بافتراض أن r_j تمثل معدل الوقوع فإن:

$$r_j = \frac{\text{عدد الأحداث}}{\text{عدد مرات احتمال وقوع الحدث}} \qquad [4.10]$$

المقام في المعادلة أعلاه [10.4] يُطلق عليه "التعرّض" وفي العـــادة يـــتم قياسه بوحدات مثل شخص/ سنة، نقوم بصياغة النموذج اللوغاريتمي لمعدل الوقوع كدالة خطيّة لواحد أو أكثر من المتغيرات التنبؤية:

$$\ln(r_t) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \qquad [a5.10]$$

وبنفس الطريقة، فإن النموذج أعلاه يشرح لوغاريتمات وحدات الأحداث المتوقعة:

$$\ln(\textit{العدد المتوقع}) = \ln(\textit{التعرّض}) + \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \qquad [b5.10]$$

وبافتراض أن التقدم في توزيع بواسون يوضح الحدث المراد دراســـته، فإن انحدار بواسون سوف يحسب تقدير الاحتمال الأعلى لمعلميات β.

سوف يكون لدينا مثال يحتوي على بيانات عن التعـــرض للإشـــعاعات والموت بمرض السرطان بين عمال المعمل الوطني في أوك ريدج بالولايات المتحدة، هناك 56 مشاهدة بملف البيانات *oakridege.dta* تمثل 56 عمر/ فئة تعرضت للإشعاع (7 فئات عمرية × 8 فئات تعرضـــت للإشـــعاع)، لكـــل

مجموعة نحن نعرف عدد حالات الوفاة، وعدد الفئات العمرية المتعرضة للإشعاع.

```
.use C:\data\oakridge.dta, clear
.describe
Contains data from C:\data\oakridge.dta
  obs:          56                          Radiation (Selvin 1995:474)
 vars:           4                          2 Jul 2012 06:11
 size:         392
--------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
--------------------------------------------------------------------------------
age             byte   %9.0g       ageg       Age group
rad             byte   %9.0g                  Radiation exposure level
deaths          byte   %9.0g                  Number of deaths
pyears          float  %9.0g                  Person-years
--------------------------------------------------------------------------------
Sorted by:
.summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
age	56	4	2.0181	1	7
rad	56	4.5	2.312024	1	8
deaths	56	1.839286	3.178203	0	16
pyears	56	3807.679	10455.91	23	71382

```
.list in 1/6
```

	age	rad	deaths	pyears
1.	< 45	1	0	29901
2.	45-49	1	1	6251
3.	50-54	1	4	5251
4.	55-59	1	3	4126
5.	60-64	1	3	2778
6.	65-69	1	1	1607

هل معدل الوفيات زاد مع التعرض للإشعاع؟ انحدار بواسون وجد أن هناك تأثيرًا ذا معنوية إحصائية:

```
. poisson deathsrad, nolog exposure(pyears) irr
```

```
Poisson regression                              Number of obs   =        56
                                                LR chi2(1)      =     14.87
                                                Prob > chi2     =    0.0001
Log likelihood =  -169.7364                     Pseudo R2       =    0.0420

------------------------------------------------------------------------------
      deaths |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         rad |   1.236469   .0603551     4.35   0.000     1.123657    1.360606
       _cons |    .000288   .0000483   -48.65   0.000     .0002074       .0004
 ln(pyears)  |          1  (exposure)
------------------------------------------------------------------------------
```

بالنسبة لتحليل الانحدار أعلاه، قمنا بتحديد عدد مرات وقــوع الحــدث (*deaths*) كمتغير تابع، والإشعاع (*rad*) كمتغير مــستقل. متغيــر التعــرّض لبواسون هو *pyears* أو شخص/ سنة في كل فئة للمتغير *rad*، الخيار **irr** يقوم بعرض نسب معدل الوقوع بدلاً من مُعَامِلات الانحدار في جــدول النتــائج، حيث حصلنا على exp(β) بدلاً من قيمة β فقط وهو الوضــع الافتراضــي. وحسب نسبة معدل الوقوع، فإن معدل الوفيات أصبح 1.236 مــرة أعلــى (زيادة بنسبة 23.6%) مع كل زيادة في فئة الإشعاع، وبالرغم من أن ذلــك المعدل ذو معنوية إحصائية، فإنه غير متناسب بدرجة كبيرة، حيــث إن R^2 الوهمية (المعادلة [9.4]) تساوي 0.042 فقط.

وللقيام باختبار حُسن المطابقة الذي يقوم بمقارنة توقعات نموذج بواسون مع الأعداد المشاهدة نقوم باستخدام أمر ما بعد التقدير **estat gof**.

```
.estat gof
 Deviance goodness-of-fit =  254.5475
 Prob > chi2(54)          =    0.0000

 Pearson goodness-of-fit  =  419.0209
 Prob > chi2(54)          =    0.0000
```

نتائج اختبار حُسن المطابقة تشير إلى أن توقعات النموذج تختلف بدرجة كبيرة عن الأعداد الحقيقية. وهذه إشارة أخرى إلى أن النموذج سئ بدرجــة كبيرة.

يمكننا الحصول على نتائج أفضل عندما نقــوم بــإدراج المتغيــر *age* كمتغير تنبؤي. حيث تزداد قيمة R^2 الوهمية لتكون 0.5966، كما أن اختبار حُسن المطابقة لا يقود إلى رفض النموذج.

```
.poisson deaths rad age, nolog exposure(pyears)
 irr

Poisson regression                                Number of obs   =         56
                                                  LR chi2(2)      =     211.41
                                                  Prob > chi2     =     0.0000
Log likelihood =   -71.4653                       Pseudo R2       =     0.5966
```

deaths	IRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
rad	1.176673	.0593446	3.23	0.001	1.065924	1.298929
age	1.960034	.0997536	13.22	0.000	1.773955	2.165631
_cons	.0000297	9.05e-06	-34.18	0.000	.0000163	.0000539
ln(pyears)	1	(exposure)				

```
.poisgof
  Deviance goodness-of-fit =  58.00534
  Prob > chi2(53)          =    0.2960

  Pearson goodness-of-fit  =  51.91816
  Prob > chi2(53)          =    0.5163
```

حتى الآن قمنا بمعاملة المتغير *rad* والمتغيــر *age* كأنهمــا متغيــران متصلان، ونحن نتوقع أن تأثيرهما على معدل الوفيات سوف يكون تــأثيرًا خطّياً، وفي الحقيقة كلا المتغيرين المستقلين تم قياسهما كفئات مرتبة، فمــثلاً عندما تكون *rad* = 1 فهذا يعني 0 تعرّض للإشعاع، *rad* = 2 يعني أنه من 0 إلى 19 مللي، *rad* = 3 يعني من 20 إلى 39 مللي وهكذا، الطريقة البديلــة لإدراج فئات التعرّض للإشعاع في تحليل الانحدار لإيجاد التــأثيرات غيــر الخطيّة يتم من خلال إدراجها كمجموعة متغيرات تنبؤية باســتخدام تــدوين المتغير العاملي ببرنامج ستاتا، المصطلح *i.rad* في النموذج أدناه يقوم بتحديد {0, 1} كمؤشر لكل فئة للمتغير *rad* تم إدراجها مع المتغيــر التنبــؤي *age*، وعندما تكون *rad* = 1 فإنه يتم إهمالها تلقائياً ويتم اعتبارها فئة أساسية.

```
.poisson deaths i.radage, nolog
 exposure(pyears) irr
```

```
Poisson regression                          Number of obs   =         56
                                            LR chi2(8)      =     215.44
                                            Prob > chi2     =     0.0000
Log likelihood = -69.451814                 Pseudo R2       =     0.6080
```

deaths	IRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
rad						
2	1.473591	.426898	1.34	0.181	.8351884	2.599975
3	1.630688	.6659257	1.20	0.231	.732428	3.630587
4	2.375967	1.088835	1.89	0.059	.9677429	5.833389
5	.7278113	.7518255	-0.31	0.758	.0961018	5.511957
6	1.168477	1.20691	0.15	0.880	.1543195	8.847472
7	4.433729	3.337738	1.98	0.048	1.013863	19.38915
8	3.89188	1.640978	3.22	0.001	1.703168	8.893267
age	1.961907	.1000652	13.21	0.000	1.775267	2.168169
_cons	.0000295	.0000106	-29.03	0.000	.0000146	.0000597
ln(pyears)	1	(exposure)				

هذا التعقيد الإضافي لنموذج المتغير التنبؤي قام بجعل النموذج أكثر ملاءمة، حيث أضاف إلى تفسير النتائج. فالتأثير الكلي للإشعاع على معدل الوفيات يبدو أنه كانت نتيجة أساسية لأعلى مستويين من مستويات الإشعاع (*rad* = 7، *rad* = 8، وهي ترتبط بـ 100 إلى 119 و120 مللي)، وعند هذه المستويات، فإن معدلات الوقوع تكون أعلى بأربع مرات تقريباً.

مستويات الإشعاع 7 و8 يبدو أن لها تأثيرات متشابهة، ولذا فإننا قد نبسّط النموذج من خلال توحيد هذه المستويات. أولاً سوف نختبر ما إذا كانت المُعَامِلات تختلف بشكل جوهري، وهي في الحقيقة لا تختلف بشكل جوهري:

```
.test 7.rad = 8.rad

 ( 1)  [deaths]7.rad - [deaths]8.rad = 0

           chi2(  1) =    0.03
         Prob > chi2 =    0.8676
```

ثم بعد ذلك نقوم بإنشاء متغير وهمي جديد باسم *rad78* وهو يساوي 1 إذا كان *rad* يساوي 7 أو 8، ونستخدم هذا المتغير الجديد بدلاً من المؤشرات *rad* = 7 و *rad* = 8، الأوامر أدناه توضح كيف يمكننا القيام بذلك في تدوين المتغير العاملي.

```
.generate rad78 = (7.rad | 8.rad)
.poisson deaths i(1/6).radrad78age, irr
 ex(pyears) nolog
```

```
Poisson regression                      Number of obs   =        56
                                        LR chi2(7)      =    215.41
                                        Prob > chi2     =    0.0000
Log likelihood = -69.465332             Pseudo R2       =    0.6079
```

deaths	IRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
rad						
2	1.473602	.4269013	1.34	0.181	.8351949	2.599996
3	1.630718	.6659381	1.20	0.231	.7324415	3.630655
4	2.376065	1.08888	1.89	0.059	.9677823	5.833629
5	.7278387	.7518538	-0.31	0.758	.0961055	5.512165
6	1.168507	1.206942	0.15	0.880	.1543236	8.847704
rad78	3.980326	1.580024	3.48	0.001	1.828214	8.665833
age	1.961722	.100043	13.21	0.000	1.775122	2.167937
_cons	.0000296	.0000106	-29.03	0.000	.0000146	.0000598
ln(pyears)	1	(exposure)				

يمكننا الاستمرار في تبسيط النموذج أكثر بنفس هذه الطريقة، وفي كل خطوة فإن الأمر test يساعدنا في تقييم ما إذا كان توحيد متغيرين وهميين يمكن تبريره.

النماذج الخطيّة العامة : Generalized Linear Models

النماذج الخطية العامة (GLM) تكون صيغتها كما يلي:

$$g[E(y)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k y \sim F \qquad [6.10]$$

حيث إن [ ]g عبارة عن دالة الربط، *F* مجموعة التوزيع، وهذه الصيغة العامة تتضمن العديد من النماذج المحددة. فعلى سبيل المثال، إذا كانت [ ]g

هي دالة الوحدة، و y تتبع توزيعاً طبيعياً (جاوس)، فإن صيغة نموذج الانحدار الخطي تكون:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k y \sim \text{توزيع طبيعي} \quad [7.10]$$

وإذا كانت [ ]g دالة لوغاريتمية، و y تتبع توزيع برنولي، فإن صيغة نموذج الانحدار اللوغاريتمي تكون:

$$\text{Logit}[E(y)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k y \sim \text{توزيع برنولي} \quad [8.10]$$

وحيث إن هذا النموذج قابل للتطبيق في مجالات عدة، فإن GLM كان يمكن استخدامه في أجزاء مختلفة من هذا الكتاب، وعلاقة GLM بهذا الفصل تأتي من قدرته على التناسب مع نماذج الأحداث. فمثلاً انحدار بواسون يتطلب أن تكون [ ]g دالة لوغاريتم طبيعي، وأن y يتبع توزيع بواسون.

$$\ln[E(y)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k y \sim \text{توزيع بواسون} \quad [9.10]$$

وكما هو متوقع مع كل طريقة مرنة، فإن الأمر **glm** ببرنامج ستاتا يسمح بالعديد من الخيارات. والمستخدمون يمكنهم تحديد ليس فقط نوع التوزيع ودالة الربط، وإنما إيضاً تفاصيل تقدير التباين، والإجراء المناسب، والمخرجات والتعرّض. هذه الخيارات تجعل الأمر **glm** بديلاً مفيدًا حتى عند تطبيقه على النماذج التي لها أوامر خاصة بها موجودة مسبقاً (مثل **regress, logistic, poisson**).

يمكننا كتابة الأمر **glm** بصيغته العامة كما يلي:

```
.glm y x1 x2 x3, family(familyname)
 link(linkname)
     exposure(expvar) eform vce(jackknife)
```

حيث إن **family( )** يحدد نوع توزيع y، **link( )** دالة الربط، **exposure( )** متغير التعرّض مثل ذلك الذي نحتاج إليه في انحدار بواسون، والخيار **eform** يقوم بعرض مُعَامِلات الانحدار في شكل أسّي حيث إنها تظهر على شكل exp(β) بدلاً من β، ويتم تقدير الأخطاء المعيارية من خلال حسابات جاكنيف jackknife calculations.

وأنواع التوزيعات المحتملة هي:

family(gaussian) توزيع جاوس أو التوزيع الطبيعي (وهذا هو الوضع الافتراضي).

family(igaussian) معكوس توزيع جاوس.

family(binomial) توزيع برنولي ذو الحدين.

family(poisson) توزيع بواسون.

family(nbinomial) التوزيع السالب ذو الحدين.

family(gamma) توزيع جاما.

يمكننا أيضاً تحديد رقم أو متغير يشير إلى كسر ذي حدين N (عدد المحاولات) أو رقم يشير إلى التباين السالب ذي حدين ودوال الانحراف، وذلك من خلال تحديدها في الخيار **family( )** كما يلي:

family(binomial#)
family(binomial *varname*)
family(nbinomial #)

دوال الربط المحتملة هي:

link(identity) دالة الربط الموحدة (هذا هو الوضع الافتراضي).

link(log) لوغاريتم.

link(logit) الدالة اللوغاريتمية.

link(probit) دالة الاحتمال.

link(cloglog) دالة المتمم اللوغاريتمي.

link(opower #) قوة الاحتمالات.

link(power #) دالة القوة.

link(nbinomial) ذو الحدين السالب.

link(loglog) لوغاريتم – لوغاريتم.

link(logc) اللوغاريتم المتمم.

مُعامِل التباين أو الأخطاء المعيارية يمكن تقديرها بعدة طرق، جزء من خيارات تقدير التباين **glm** هي كما يلي:

opg مقدّر التباين لـ هال Hall وهال وهوسمان بي إتش التكعيبي Hall and Hausman B-H-cubed، وبرندت Berndt.

oim مقدّر تباين مصفوفة المعلومات المشاهدة.

robust مقدّر الشطيرة لتباين هوبر/وايت Huber/White.

unbiased مقدّر الشطيرة غير المتحيز للتباين.

nwest مقدّر تباين الارتباط الذاتي الثابت واختلاف التباين.

jackknife تقدير جاكنيف للتباين.

jackknife1 تقدير جاكنيف ذو الخطوة الواحدة للتباين.

bootstrap تقدير بوتستراب Bootstrap للتباين، الوضع الافتراضي هو 199 تكرارًا، ويمكن تحديد رقم تكرار معين عن طريق إضافة ذلك من خلال الخيار **bstrep(#)**

للحصول على قائمة كاملة بالخيارات مع بعض التفاصيل التقنية حولها، قم بالبحث عن **glm** في دليل المستخدم *Base Reference Manual*، وللحصول على شرح أكثر تفصيلاً عن موضوعات GLM يمكنك الاطلاع على دراسة Hardin and Hilbe (2012).

الفصل (7) بدأ مع الانحدار البسيط لمتوسط العمر المتوقع على متوسط سنوات الدراسة في 188 دولة.

```
.use C:\data\Nations2.dta, clear
.regress life school
```

```
      Source |       SS       df       MS              Number of obs =     188
-------------+------------------------------           F(  1,   186) =  206.34
       Model |  9846.65406     1  9846.65406           Prob > F      =  0.0000
    Residual |  8875.86926   186  47.7197272           R-squared     =  0.5259
-------------+------------------------------           Adj R-squared =  0.5234
       Total |  18722.5233   187  100.120446           Root MSE      =  6.9079

------------------------------------------------------------------------------
        life |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      school |    2.45184   .1706856    14.36   0.000     2.115112    2.788569
       _cons |   50.35941    1.36924    36.78   0.000     47.65817    53.06065
------------------------------------------------------------------------------
```

يمكننا إنشاء نفس النموذج والحصول على نفس التقديرات من خلال استخدام الأمر **glm**.

.glm *life school,* link(identity) family(gaussian)

```
Iteration 0:   log likelihood = -629.09751

Generalized linear models                          No. of obs      =       188
Optimization     : ML                              Residual df     =       186
                                                   Scale parameter =  47.71973
Deviance         =  8875.869256                    (1/df) Deviance =  47.71973
Pearson          =  8875.869256                    (1/df) Pearson  =  47.71973

Variance function: V(u) = 1                        [Gaussian]
Link function    : g(u) = u                        [Identity]

                                                   AIC             =  6.713803
Log likelihood   = -629.0975058                    BIC             =  7901.891

------------------------------------------------------------------------------
             |                 OIM
        life |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      school |    2.45184   .1706856    14.36   0.000     2.117303    2.786378
       _cons |   50.35941    1.36924    36.78   0.000     47.67575    53.04307
------------------------------------------------------------------------------
```

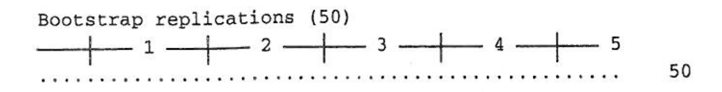
وحيث إن الخيار **link(identity)** والخيار **family(gaussian)** هما الخيارات الافتراضية، فيمكننا تركهما وعدم طباعتهما في الأمر أعلاه **glm**، وسوف نحصل على نفس النتائج.

كما أنه بالإمكان صياغة نفس النموذج OLS ولكن نحصل على أخطـــاء معيارية مقدّرة عن طريق بوتستراب.

```
.glm life school, link(identity)
 family(gaussian) vce(bootstrap)

(running glm on estimation sample)

Bootstrap replications (50)
```



```
Generalized linear models                   No. of obs      =       188
Optimization     : ML                       Residual df     =       186
                                            Scale parameter =  47.71973
Deviance         =  8875.869256             (1/df) Deviance =  47.71973
Pearson          =  8875.869256             (1/df) Pearson  =  47.71973

Variance function: V(u) = 1                 [Gaussian]
Link function    : g(u) = u                 [Identity]

                                            AIC             =  6.713803
Log likelihood   = -629.0975058             BIC             =  7901.891
```

life	Observed Coef.	Bootstrap Std. Err.	z	P>|z|	Normal-based [95% Conf.	Interval]
school	2.45184	.1436837	17.06	0.000	2.170225	2.733455
_cons	50.35941	1.29733	38.82	0.000	47.81669	52.90213

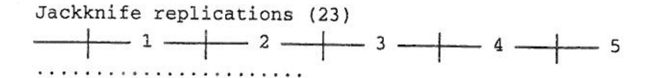
الأخطاء المعيارية لبوتستراب تعكس التباين المشاهد بـين المُعَـاملات المقدّرة من 50 عينة تحتوي على 188 = *n*. حالة تم الحصول عليها بطريقة المعاينة العشوائية مع استبدال البيانات الأصلية 188 = *n*. فـي هـذا المثـال الأخطاء المعيارية لبوتستراب أقل من الأخطـاء المعياريـة النظريـة ذات العلاقة ونتج عنها فترات ثقة أصغر.

وبالمثل، فإنه يمكننا استخدام الأمر **glm** لتكرار الانحدار اللوغـاريتمي مع بيانات مكوك الفضاء الذي سبق لنا استخدامها في الفصل (9). في هـذا المثال، قمنا بحساب الأخطاء المعيارية لجـاكنيف ومعـدل الاحتمـالات أو عرض المُعَامِلات في شكل أسّي **(eform)**

```
.use C:\data\shuttle0.dta, clear
.glm any date, link(logit) family(bernoulli)
 eform vce(jackknife)
```

```
(running glm on estimation sample)

Jackknife replications (23)
```



```
Generalized linear models                    No. of obs      =         23
Optimization     : ML                        Residual df     =         21
                                             Scale parameter =          1
Deviance         =  25.98219269              (1/df) Deviance =   1.237247
Pearson          =   22.8885488              (1/df) Pearson  =   1.089931

Variance function: V(u) = u*(1-u/1)          [Binomial]
Link function    : g(u) = ln(u/(1-u))        [Logit]

                                             AIC             =   1.303574
Log likelihood   = -12.99109634              BIC             =  -39.86319
```

any	Odds Ratio	Jackknife Std. Err.	t	P>|t|	[95% Conf.	Interval]
date	1.002093	.0015797	1.33	0.198	.9988222	1.005374
_cons	1.34e-08	1.89e-07	-1.28	0.214	2.32e-21	76840.52

انحدار بواسون مع متغيرات تنبؤية تم التطرق إليه سابقاً في هذا الفصل.

```
.use C:\data\oakridge.dta, clear
.poisson deaths i.rad age, nolog exposure(pyears)
 irr
```

وهو يتوافق مع نموذج glm أدناه:

```
.glm deaths i.rad age, link(log)
 family(poisson) exposure(pyears) eform
```

```
Iteration 0:   log likelihood =  -75.68551
Iteration 1:   log likelihood = -69.595462
Iteration 2:   log likelihood = -69.452909
Iteration 3:   log likelihood = -69.451814
Iteration 4:   log likelihood = -69.451814

Generalized linear models                          No. of obs      =        56
Optimization     : ML                              Residual df     =        47
                                                   Scale parameter =         1
Deviance         =  53.97836926                    (1/df) Deviance =  1.148476
Pearson          =  53.59824023                    (1/df) Pearson  =  1.140388

Variance function: V(u) = u                        [Poisson]
Link function    : g(u) = ln(u)                    [Log]

                                                   AIC             =  2.801851
Log likelihood   =   -69.451814                    BIC             = -135.2132
```

deaths	IRR	OIM Std. Err.	z	P>|z|	[95% Conf.	Interval]
rad						
2	1.473591	.426898	1.34	0.181	.8351884	2.599975
3	1.630688	.6659257	1.20	0.231	.732428	3.630587
4	2.375967	1.088835	1.89	0.059	.9677429	5.833389
5	.7278114	.7518256	-0.31	0.758	.0961019	5.511958
6	1.168477	1.20691	0.15	0.880	.1543196	8.847472
7	4.433727	3.337737	1.98	0.048	1.013862	19.38915
8	3.89188	1.640978	3.22	0.001	1.703168	8.893267
age	1.961907	.1000652	13.21	0.000	1.775267	2.168169
_cons	.0000295	.0000106	-29.03	0.000	.0000146	.0000597
ln(pyears)	1	(exposure)				

بالرغم من أن الأمر **glm** يكرر النماذج التي تم صياغتها بواسطة أوامر متخصصة، كما أضاف بعض القدرات الجديدة. هذه الأوامر المتخصصة لها ميزاتها التي منها السرعة وتوافر خيارات التخصيص. الميزة الأساسية للأمر **glm** هي قدرته على التلاؤم مع نماذج ستاتا التي ليس لها أوامر خاصة بها.

الفصل الحادي عشر

تحليل المكوّنات الرئيسية
التحليل العَاملِي والتحليل العنقودي
Principal Component, Factor And Cluster Analysis

تحليل المكونات الرئيسة، والتحليل العاملي تعتبر طرق تبسيط وتوحيد العديد من المتغيرات المترابطة في عدد أصغر للأبعاد الضمنية. خلال خطوات التبسيط، يجب على المحلل الاختيار من بين عدد كبير من الخيارات الصعبة. وإذا كانت البيانات تعكس أبعادًا ضمنية مختلفة، فإن الخيارات المختلفة قد تؤدي إلى الحصول على نتائج متشابهة. وفي غياب أبعاد ضمنية مختلفة، فإن الخيارات المختلفة قد تؤدي إلى الحصول على نتائج مختلفة. التحليل باستخدام هذه الخيارات يوضح كيف أن نتيجة معينة مستقرة أو إلى أي مدى تعتمد هذه النتيجة على خيارات عشوائية يتم اختيارها بناءً على الطريقة المستخدمة في التحليل.

يقوم برنامج ستاتا بإجراء تحليل المكونات الرئيسة، والتحليل العاملي مستخدماً خمسة أوامر رئيسة هي:

Pca تحليل المكونات الرئيسة.

Factor استخراج عوامل وأنواع مختلفة متعددة تتضمن المكونات الرئيسة.

Screeplot يقوم بإنشاء رسم الحصى والحجارة (رسم بياني للجذر الكامن) من آخر نتيجة للأمر **pca** أو الأمر **factor**.

Rotate يقوم بحساب التدوير المتعامد (عوامل غير مترابطة) أو التدوير المائل (عوامل مترابطة) بعد الأمر **factor**.

Predict يقوم بإنشاء درجات عوامل (متغيرات مركبة) وإحصائيات لحالات أخرى بعد الأمر **factor** أو الأمر **pca** أو الأمر **rotate**.

درجات العوامل أو المتغيرات المركبة يتم إنشاؤها عن طريق الأمر **predict**، ويمكن حفظها ووضعها في قوائم وتمثيلها بيانياً وتحليلها مثلها مثل متغيرات ستاتا الأخرى، سوف يتم توضيح مثل هذا التحليل في جزء جديد باستخدام مثال يتضمن بيانات دراسة استقصائية.

المستخدمون الذين يقومون بإنشاء متغيرات مركبة باستخدام طرق معقدة، تقوم بإضافة متغيرات أخرى معاً بدون القيام بإجراء تحليل عاملي يمكنهم من تقييم نتائجهم من خلال احتساب α (ألفا) معامل الثبات:

ألفا ثبات كرونباخ α Cronbach.

فبدلاً من دمج المتغيرات، فإن التحليل العنقودي يدمج المشاهدات، وذلك من خلال إيجاد عدم التداخل، وإنشاء مجموعات أو تصنيفات على أساس عملي. طرق التحليل العنقودي أكثر تنوعاً من طرق التحليل العاملي. والأمر **cluster** له عدة أدوات يمكنه من خلالها إجراء التحليل العنقودي، وتمثيل النتائج بيانياً، وإنشاء متغيرات جديدة لتحديد المجموعات التي ظهرت في النتائج. تحليل المكونات الرئيسة، والتحليل العاملي، والتحليل العنقودي، والأوامر المتعلقة بها تم توضيحها بالتفصيل في دليل المستخدم الخاص ببرنامج ستاتا *Multivariate Statistics Reference Manual*.

يختم هذا الفصل موضوعه بإلقاء نظرة ثانية على قدرات نماذج المعادلة المركبة لبرنامج ستاتا (**sem**) المستخدمة لقياس النماذج التي تتضمن قياس نموذج المكونات.

الطرق التي سوف يتم شرحها في هذا الفصل، يمكن الوصول إليها عبر قوائم ستاتا التالية:

Statistics > Multivariate analysis
Graphics > Multivariate analysis graphs
Statistics > SEM (structural equation modeling)

أمثلة عن الأوامر : Example Commands

```
.pca x1-x20
```

يقوم بحساب المكونات الرئيسة للمتغيرات التي تبدأ من المتغير *x1* وحتى المتغير *x20*.

```
.pca x1-x20, mineigen(1)
```

يقوم بحساب المكونات الرئيسة للمتغيرات التي تبدأ من المتغير *x1* وحتى المتغير *x20*، ويقوم بحفظ المكونات التي تكون قيمة الجذر الكامن لها أكبر من 1.

```
.factor x1-x20, ml factor(5)
```

يقوم بحساب التحليل العاملي بطريقة الأرجحية العظمى للمتغيرات من *x1* وحتى *x2*، ويقوم بحفظ أول خمسة عوامل فقط.

```
.screeplot
```

يقوم بإنشاء رسم بياني للحصى والحجارة للجذر الكامن لعدد من المكونات تبدأ من آخر نتائج للأمر **factor**.

```
.rotate, varimax factors(2)
```

يقوم بحساب التدوير المتعامد (أكبر تباين) لأول عاملين اثنين من آخر نتائج للأمر **factor**.

```
.rotate, promax factors(3)
```

يقوم بحساب التدوير المائل (تدوير المحاور) لأول ثلاثة عوامل من آخر نتائج للأمر **factor**.

```
.predict f1 f2 f3
```

يقوم هذا الأمر بإنشاء ثلاثة متغيرات عامليّة جديدة باسم *f3, f2, f1* بناءً على آخر نتائج للأمر **factor** وللأمر **rotate**.

```
.alpha x1-x10
```

يقوم بحساب معدل الثبات α لكرونباخ لمتغير مركب ويُعرّف بأنه مجموع المتغيرات من *x1* إلى *x10*، وعند الإدخال سوف يتم عكس القيم السلبية. والخيارات التي تتعلق بهذا الأمر يمكنها تغيير هذا الوضع الافتراضي، أو إنشاء متغير تركيبي من خلال إضافة المتغيرات الأصلية أو قيمها المعيارية.

```
.cluster centroidlinkage x y z w, measure(L2)
 name(L2cent)
```

يقوم هذا الأمر بحساب التحليل العنقودي التراكمي مع ارتباط لنقطة التقاطع باستخدام المتغيرات *x, y, z, w*، وتقوم مسافة إقليدس (L2) بقياس الاختلاف بين المشاهدات. والنتائج من هذا التحليل العنقودي يتم حفظها باسم *L2cent*.

```
.cluster dendrogram, ylabel(0(.5)3) cutnumber(20)
     xlabel(, angle(vertical))
```

يقوم بإنشاء رسم بياني لشجرة التحليل العنقودي تعرض نتائج آخر أمر للتحليل العنقودي. الخيار **cutnumber(20)** يحدد بأن الرسم البياني يبدأ مع الـ 20 عنقودًا المتبقية بعد اضمحلال أغلب المشاهدات المتشابهة. وتتم طباعة توصيفات لهذا الشكل البياني في شكل عمودي أسفل الشكل نفسه.

```
.cluster generate ctype = groups(3), name(L2cent)
```

يقوم بإنشاء متغير جديد باسم *ctype* (قيمه 1 أو 2 أو 3) تقوم بتصنيف كل مشاهدة في واحدة من أعلى ثلاث مجموعات تم إنشاؤها باستخدام التحليل العنقودي الذي توجد نتائجه في *L2cent*.

تحليل المكونات الرئيسة والتحليل العاملي للمكونات الرئيسة :

Principal Component Analysis and Principal Component Factoring

لتوضيح التحليل العاملي، وتحليل المكونات الرئيسة، سوف نبدأ مع مجموعة بيانات صغيرة موجودة بالملف *planets.dta* تشرح التسع كواكب التقليدية بالنظام الشمسي (هذه البيانات من دراسة Beatty et al. 1981)،

البيانات تتضمن مجموعة من المتغيرات في شكل خام وشكل لوغــاريتمي. وتم استخدام اللوغاريتمات هنا لتقليل الالتواء، وجعل العلاقات أكثر خطيّــة بين المتغيرات.

```
.use C:\data\planets.dta, clear
.describe
```

```
Contains data from C:\data\planets.dta
  obs:            9                          Solar system data
 vars:           12                          2 Jul 2012 06:11
 size:          405
```

variable name	storage type	display format	value label	variable label
planet	str7	%9s		Planet
dsun	float	%9.0g		Mean dist. sun, km*10^6
radius	float	%9.0g		Equatorial radius in km
rings	byte	%8.0g	ringlbl	Has rings?
moons	byte	%8.0g		Number of known moons
mass	float	%9.0g		Mass in kilograms
density	float	%9.0g		Mean density, g/cm^3
logdsun	float	%9.0g		natural log dsun
lograd	float	%9.0g		natural log radius
logmoons	float	%9.0g		natural log (moons + 1)
logmass	float	%9.0g		natural log mass
logdense	float	%9.0g		natural log dense

```
Sorted by:  dsun
```

نتائج تحليل المكونات الرئيسة توضح بأن هناك تركيباً خطيــاً يــشرح أعلى قيمة للتباين في المتغيرات التي تم مشاهداتها ويسمى "مكــوّن رئــيس أول". كما أن النتائج أوضحت بأن هناك تركيباً خطّياً متعامدًا آخــر (غيــر مترابط) يشرح أعلى قيمة تباين متبقية "مكوّن رئيس ثاني" وهكذا حتى يــتم شرح كل قيم التباين. من المتغيرات k يمكننا استخراج المكونات الرئيسة لـــ k والتي يمكنها شرح كل قيم التباين. تحليل المكونات الرئيسة يمكن استخدامه كأداة لاختزال البيانات، لأن استخدام مكونات k أقل سوف يشرح جزءًا كبيرًا من التباين، وإذا تم تركيز العمل بشكل أكبر على المكونات، فإنه بالإمكــان تبسيط التحليل.

وعند تطبيق تحليل المكونات الرئيسة علـــى ســتة متغيــرات توضـــح الكواكب، فإننا نحصل على مكونات رئيسة تشرح التباين بالكامل:

```
.pca ringslogdsun-logdense
```

```
Principal components/correlation                 Number of obs    =          9
                                                 Number of comp.  =          6
                                                 Trace            =          6
    Rotation: (unrotated = principal)            Rho              =     1.0000
```

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	4.62365	3.45469	0.7706	0.7706
Comp2	1.16896	1.05664	0.1948	0.9654
Comp3	.112323	.0539515	0.0187	0.9842
Comp4	.0583717	.0217421	0.0097	0.9939
Comp5	.0366296	.0365651	0.0061	1.0000
Comp6	.00006454	.	0.0000	1.0000

Principal components (eigenvectors)

Variable	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Unexplained
rings	0.4554	0.0714	0.2912	0.0351	-0.8370	0.0301	0
logdsun	0.3121	-0.6576	0.5930	-0.1418	0.3135	-0.0156	0
lograd	0.4292	0.3455	-0.0390	-0.3216	0.2619	0.7231	0
logmoons	0.4541	0.0003	-0.1567	0.8466	0.2286	0.0156	0
logmass	0.3878	0.5037	0.1374	-0.2427	0.2675	-0.6682	0
logdense	-0.3930	0.4352	0.7201	0.3157	0.0932	0.1708	0

نتائج الأمر **pca** توضح لنا أن أول مكونين اثنين يشرحان أكثـر مــن 96% من التباين التراكمي للستة متغيرات بالكامل. القيم الكامنـــة المتعلقـــة بالتباين المعياري تم شرحها بواسطة كل مكوّن. مجموع التباين المعيـــاري للمتغيرات الستة بالكامل هو 6. ومن هذا المجموع نرى أن المكـــوّن الأول Comp1 يشرح 4.62365 والذي تتم مقارنته مـــع العـــدد الكلـــي للمكوّنـــات 4.62365 ÷ 6 = 0.7706 أو حوالي 77% من المجمـــوع الكلـــي، المكـــوّن الثاني Comp2 يشرح 1.16896 ÷ 6 = 0.1948 أو حوالي 19% إضـــافية، تحليل المكوّنات له قيم كامنة أقل من 1.0 وهذه القيم تشرح أقل من مكـــافئ تباين متغير واحد وهذا لا يساعدنا في اختزال البيانات. محللو البيانات فـــي العادة يستبعدون المكونات الثانوية، ويركزون على المكونات التي لهـــا قـــيم كامنة تساوي 1 على الأقل.

وأفضل طريقة لإجراء اختزال للبيانات تتم من خلال الأمر **factor** حيث يمكن استخدام العديد من الخيارات مع هذا الأمر، وهي تتعلق بتحليل التحليل العاملي للمكونات الرئيسة. وللحصول على عوامل المكونات الرئيسة نقــوم بطباعة الأمر التالي:

```
.factor rings logdsun - logdense, pcf
(obs=9)

Factor analysis/correlation                    Number of obs    =      9
    Method: principal-component factors        Retained factors =      2
    Rotation: (unrotated)                      Number of params =     11
```

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	4.62365	3.45469	0.7706	0.7706
Factor2	1.16896	1.05664	0.1948	0.9654
Factor3	0.11232	0.05395	0.0187	0.9842
Factor4	0.05837	0.02174	0.0097	0.9939
Factor5	0.03663	0.03657	0.0061	1.0000
Factor6	0.00006	.	0.0000	1.0000

```
LR test: independent vs. saturated:  chi2(15) =  100.49 Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances
```

Variable	Factor1	Factor2	Uniqueness
rings	0.9792	0.0772	0.0353
logdsun	0.6710	-0.7109	0.0443
lograd	0.9229	0.3736	0.0088
logmoons	0.9765	0.0003	0.0465
logmass	0.8338	0.5446	0.0082
logdense	-0.8451	0.4705	0.0644

التحليل العاملي للمكونات الرئيسة يبدأ باستخراج المكونات الرئيسة، ثم الإبقاء على المكونات التي تفي بمعيار الأهمية وهذا يتم افتراضياً، حيث يتم الإبقاء على تلك المكونات التي تكون قيمها الكامنة أكبر من 1. كمــا رأينــا سابقاً في مثال **pca** أن أول مكونين فقط تقابل معيار الأهمية وهذان المكونان يشرحان أكثر من 96% للتباين المشترك للستة متغيرات معاً، وبذلك يمكننــا إهمال بقية المكونات وهي من المكوّن الثالث وحتى المكوّن السادس.

هناك خياران من خيارات الأمر **factor** يمكنهما التحكم في عدد العوامل المستخرجة:

factors(#) حيث إن # تحدد عدد العوامل.

mineigen(#) حيث إن # تحدد أقل قيمة كامنة للعوامل المحتفظ بها.

وبما أن التحليل العاملي للمكونات الرئيسة يستبعد بشكل تلقائي العوامل التي تقل قيمها الكامنة عن 1.

```
.factor ringslogdsun - logdense, pcf
```

وهذا مكافئ للأمر:

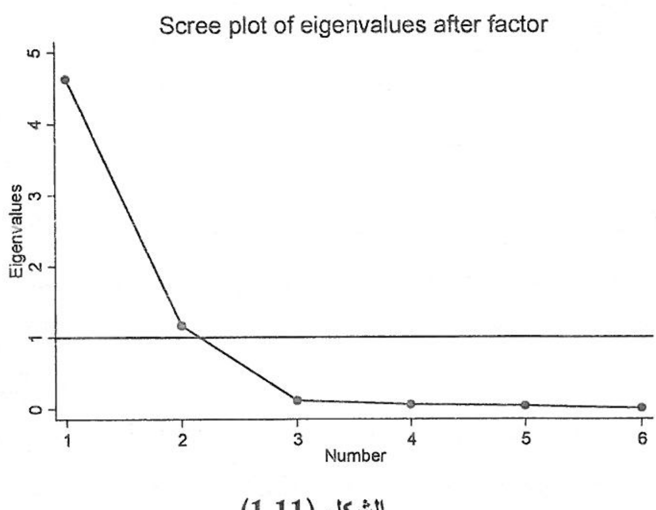
```
.factor rings logdsun - logdense, pcf mineigen(1)
```

في هذا المثال، يمكننا أن نحصل على نفس النتائج، وذلك من خلال طباعة الأمر:

```
.factor rings logdsun - logdense, pcf factors(2)
```

ولعرض نافذة الرسم البياني (الرسم البياني للقيم الكامنة مقابل عدد العوامل أو عدد المكونات) بعد أي أمر **factor** نقوم باستخدام الأمر **screeplot**، في الشكل (1.11) نرى خطاً أفقياً عند القيمة الكامنة = 1 صانعاً القطع المعتاد للمكونات الرئيسة المحتفظ بها، ومرة أخرى، فإن هذا القطع يؤكد عدم أهمية المكونات من 3 إلى 6.

```
.screeplot, yline(1)
```



الشكل (1.11)

التدوير : Rotation

التدوير يبسط أكثر تركيبة العوامل، فبعد التحليل العاملي قم بطباعة الأمر **rotate** ثم اتبع هذا الأمر بخيار يُحدد نوع التدوير. هناك نوعان شائعان للتدوير هما:

varimax التدوير المتعامد لأكبر تباين، منتجاً عوامل أو مكونات غير مترابطة (هذا هو الوضع الافتراضي).

promax() التدوير المائل بروماكس وهذا النوع يسمح بوجود العوامل أو المكونات المترابطة، اختر عدد (قوة بروماكس) أقل من أو تساوي 4، وكلما زاد هذا العدد كلما كانت درجة الارتباط العاملي أقوى، والوضع الافتراضي أن يكون **promax(3)**

وللحصول على قائمة كاملة بطرق التدوير والخيارات الأخرى قم بطباعة الأمر **help rotate**، فعلى سبيل المثال:

factors() هذا الخيار يحدد عدد العوامل التي يتم حفظها.

entropy التدويل المتعامد الأقل تنظيماً.

التدوير يمكن القيام به بعد أي تحليل عاملي وليس فقط مع التحليل العاملي للمكونات الرئيسة كما سنرى لاحقاً، هذا الجزء سوف يستخدم الأمثلة التي تم استخدامها مع الأمر **factor** والأمر **pcf**، التدوير المتعامد (الافتراضي) لأول مكوّنين تم إيجادهما في بيانات المجموعة الشمسية، وللقيام بعملية التدوير قم بطباعة **rotate**

```
.rotate
```

```
Factor analysis/correlation                     Number of obs    =        9
    Method: principal-component factors         Retained factors =        2
    Rotation: orthogonal varimax (Kaiser off)   Number of params =       11
```

Factor	Variance	Difference	Proportion	Cumulative
Factor1	3.36900	0.94539	0.5615	0.5615
Factor2	2.42361	.	0.4039	0.9654

```
LR test: independent vs. saturated:  chi2(15) =  100.49 Prob>chi2 = 0.0000
```

Rotated factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Uniqueness
rings	0.8279	0.5285	0.0353
logdsun	0.1071	0.9717	0.0443
lograd	0.9616	0.2580	0.0088
logmoons	0.7794	0.5882	0.0465
logmass	0.9936	0.0678	0.0082
logdense	-0.3909	-0.8848	0.0644

Factor rotation matrix

	Factor1	Factor2
Factor1	0.7980	
Factor2	0.6026	-0.7980

المثال أعلاه يقبل كل الأوضاع الافتراضية: التدوير المتعامد لأعلى تباين، ونفس عدد العوامل يتم حفظها من آخر أمر **factor**؛ كما يمكننا الحصول على نفس النتائج بالضبط من خلال إضافة خيارات إلى الأوامر **.rotate, varimax factors(2)**

ولإجراء تدوير بروماكس (promax) المائل (يسمح بالعوامل المترابطة) لآخر نتائج للتحليل العاملي نقوم بطباعة:

```
.rotate, promax
```

```
Factor analysis/correlation                        Number of obs    =        9
    Method: principal-component factors            Retained factors =        2
    Rotation: oblique promax (Kaiser off)          Number of params =       11
```

Factor	Variance	Proportion	Rotated factors are correlated
Factor1	4.12467	0.6874	
Factor2	3.32370	0.5539	

```
LR test: independent vs. saturated:  chi2(15) =  100.49 Prob>chi2 = 0.0000
```

Rotated factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Uniqueness
rings	0.7626	0.3466	0.0353
logdsun	-0.1727	1.0520	0.0443
lograd	0.9926	0.0060	0.0088
logmoons	0.6907	0.4275	0.0465
logmass	1.0853	-0.2154	0.0082
logdense	-0.1692	-0.8719	0.0644

Factor rotation matrix

	Factor1	Factor2
Factor1	0.9250	0.7898
Factor2	0.3800	-0.6134

هذا المثال يستخدم بشكل افتراضي قوة بروماكس 3، كما يمكننا تحديد قوة بروماكس، وتحديد عدد العوامل بالضبط كما يلي:

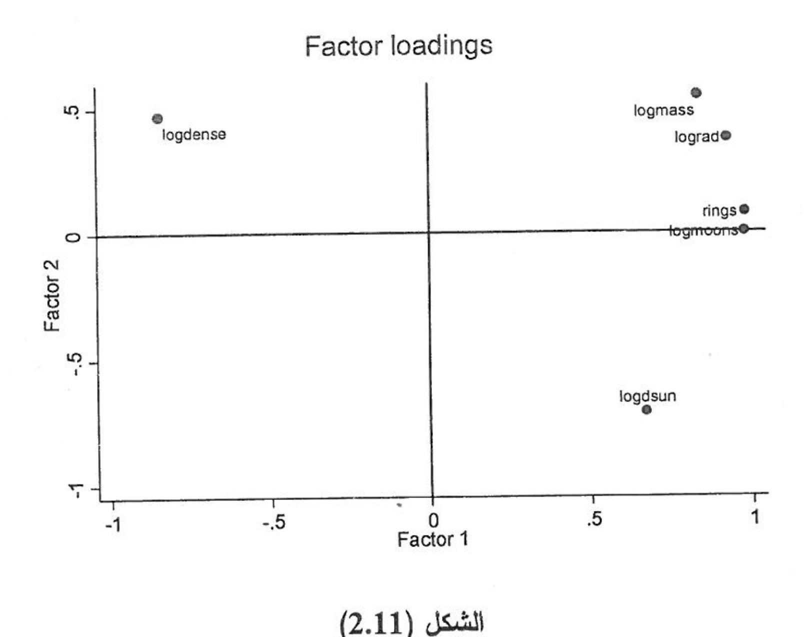
.rotate, promax(3) factors(2)

عند استخدام **promax(4)** فإن هذا سوف يسمح بإجراء تبسيط أكثر لمصفوفة التشبع، وهذا يؤدي إلى ارتباط أكثر قوة بين العوامل وأقل تفسيراً للتباين الكلي.

وبعد إجراء تدوير بروماكس المائل، فإن المتغيرات *rings, lograd, logmoons, logmass* سوف يكون الحمل أكثر ثقلاً على العامل الأول، هذا يبدو أنه ذو "حجم أكبر/ العديد من الأقمار الصناعية" "large size/many"

"satellites"، المتغير *logdsum* والمتغير *logdense* تُحمّل أكثر على العامل الثاني مما يجعله "بعيد/ أقل كثافة" "far out/low density"، كما أن الأمر **loadingplot** يقوم بإنشاء رسم بياني بعد إجراء التحليل العاملي، وهذا الرسم يساعد في تمثيل هذه النتائج بوضوح (الشكل 2.11).

```
.loadingplot, factors(2) yline(0) xline(0)
```



الشكل (2.11)

القيم العاملية : Factor Scores

القيم العاملية هي مكوّنات خطيّة يتم إنشاؤها بمعايرة كل متغير إلى متوسط صفر، وتباين الوحدة ثم وزنها مع مُعَامِلات قيم العامل وجمعها لكل عامل. الأمر **predict** يقوم بهذه الحسابات بشكل تلقائي مستخدماً آخر نتائج للأمر **rotate** أو الأمر **factor**، وعند استخدام الأمر **predict** فإننا يجب أن نقوم بإدراج أسماء المتغيرات الجديدة بعده وهذه الأسماء مثل *f1* و *f2*.

```
.predict f1 f2
```

```
(regression scoring assumed)

Scoring coefficients (method = regression)
```

Variable	Factor1	Factor2
rings	0.21177	0.06605
logdsun	0.14513	-0.60818
lograd	0.19960	0.31958
logmoons	0.21119	0.00024
logmass	0.18033	0.46591
logdense	-0.18278	0.40252

```
.label variable f1 "Large size/many satellites"
.label variable f2 "Far out/low density"
.list planet f1 f2
```

	planet	f1	f2
1.	Mercury	-.9172388	-1.256881
2.	Venus	-.5160229	-1.188757
3.	Earth	-.3939372	-1.035242
4.	Mars	-.6799535	-.5970106
5.	Jupiter	1.342658	.3841085
6.	Saturn	1.184475	.9259058
7.	Uranus	.7682409	.9347457
8.	Neptune	.647119	.8161058
9.	Pluto	-1.43534	1.017025

عند تحويل القيم إلى قيم معيارية، فإن القيم العاملية *f1* و*f2* سوف يكون لها متوسطات (تقريباً) تساوي صفر، وانحراف معياري يساوي 1.

```
.summarize f1 f2
```

Variable	Obs	Mean	Std. Dev.	Min	Max
f1	9	-3.31e-09	1	-1.43534	1.342658
f2	9	9.93e-09	1	-1.256881	1.017025

ولذلك فإن القيم العاملية يتم قياسها بوحدات انحرافاتها المعيارية عن متوسطاتها، فمثلاً في الجدول ما قبل السابق كوكب عُطارد Mercury له انحراف معياري قدره 0.92 تقريباً أقل من المتوسط "حجم كبير/ العديد من الأقمار الصناعية" (*f1*) لأن هذا الكوكب صغير وليست له أقمار صناعية، عُطارد له انحراف معياري أقل من المتوسط "بعيد جداً/ أقل كثافة" (*f2*) لأنه أقرب للشمس

وذو كثافة عالية، وعلى العكس من ذلك، فإن كوكب زُحل Saturn له انحرافات معيارية 1.18 و0.93 أعلى من المتوسط لهذين البُعدين.

تدوير بروماكس المائل يسمح بالترابط بين القيم العاملية:

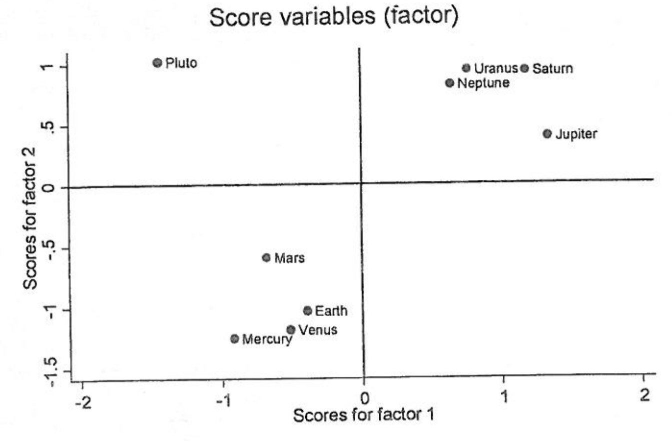
```
.correlate f1 f2
(obs=9)
```

	f1	f2
f1	1.0000	
f2	0.4974	1.0000

القيم العاملية للعامل 1 لها ارتباط موجب متوسط مع القيم العاملية للعامل 2، هذه الكواكب بعيدة جداً/ وأقل كثافة هي كواكب من المحتمل أن تكون أكبر حجماً مع العديد من الأقمار الصناعية.

الأمر الآخر الذي يتم استخدامه لإنشاء رسم بياني بعد إجراء عملية التحليل العاملي هو الأمر **scoreplot** حيث يقوم برسم شكل الانتشار لمشاهدات القيم العاملية، كما يمكن استخدام هذا الأمر مع عوامل المكونات الرئيسة، وتساعد هذا الأشكال في تحديد القيم المتطرفة المتعددة أو التحليل العنقودي للمشاهدات التي تظهر مختلفة عن البقية، الشكل (3.11) يوضح ثلاثة أنواع مختلفة من الكواكب.

```
.scoreplot, mlabel(planet) yline(0) xline(0)
```



الشكل (3.11)

الكواكب الصخرية الداخلية (مثل عُطارد Mercury، انخفاض في "الحجم الكبير/ العديد من الأقمار الصناعية" العامل 1، وانخفاض كذلك في "البعيد جداً/ منخفض الكثافة" العامل 2) تظهر معاً في الجانب الأسفل الأيسر في الشكل أعلاه، أما الكواكب العملاقة الغازية فلها خصائص عكسية تماماً، وتظهر معاً في الجانب الأعلى الأيمن، كوكب بلوتو Pluto الذي يشبه شكله بعض الأقمار الخارجية المنتظمة فهذا الكوكب فريد من نوعه مقارنة بالتسعة أقمار التقليدية، لأن بُعده "بعيد جداً/ أقل كثافة" وفي نفس الوقت فهو منخفض في بُعد "الحجم الأكبر/ العديد من الأقمار الصناعية"، لذلك فإن التحليل العاملي قام بتصنيف كوكب بلوتو كنوع آخر من الأنواع التي لا تتناسب مع المجموعتين الرئيستين من الكواكب، وبالأخذ في الاعتبار الطبيعة الخاصة لكوكب بلوتو، فإن الاتحاد الفلكي العالمي قام في 2006 بإعادة تصنيف كوكب بلوتو من كونه أحد الكواكب الرئيسة إلى واحد من الكواكب التي تُعرف باسم "الكواكب القزمية" وهذا يترك ثماني كواكب فقط.

إذا قمنا باستخدام تدوير أعلى تباين بدلاً من تدوير بروماكس، فإن القيم العاملية غير المترابطة تكون:

```
.quietly factor rings logdsun - logdense, pcf
.quietly rotate
.quietly predict varimax1 varimax2
.correlate varimax1 varimax2
```

(obs=9)

	varimax1	varimax2
varimax1	1.0000	
varimax2	0.0000	1.0000

عند إنشاء القيم العاملية باستخدام الأمر **predict**، فإن هذه القيم يمكن معاملتها مثل أي متغير من متغيرات ستاتا وإدراجها في أي أمر، وتحليل ارتباطها وتمثيلها بيانياً وهكذا، القيم العاملية في العادة تستخدم في العلوم الاجتماعية والسلوكية وذلك لتوحيد العديد من الاختبارات أو عناصر استمارات الاستبيان في متغيرات مركبة أو مؤشرات، كما سيتم شرحه لاحقاً في هذا الفصل. أما في مجال العلوم التطبيقية مثل علوم المناخ أو الاستشعار عن بُعد، فإن القيم العاملية التي يتم الحصول عليها بواسطة تحليل المكونات

الرئيسة بدون تدوير تساعد في تحليل حجم كبير من البيانات، في هذه التطبيقات فإن تحليل المكونات يُطلق عليه "الدوال المتعامدة التجريبية"، أول دالة متعامدة تجريبية – أو اختصاراً EOF1 – تساوي قيمة عاملية لأول مكوّن رئيس غير مدوّر، EOF2 هي قيمة المكون الرئيس الثاني وهكذا.

التحليل العاملي الرئيس : Principal Factoring

الأمثلة أعلاه تضمنت تحليل المكونات الرئيسة وتم استخدام الأمر **factor** مع الخيار **pcf**، أما الخيارات الأخرى للأمر **factor** فتقوم بالتحليل العاملي بطرق مختلفة.

pcf تحليل المكونات الرئيسة.

pf التحليل العاملي الرئيس (وهو الوضع الافتراضي).

ipf التحليل العاملي للمكونات مع قيم الشيوع المتكررة iterated communalities.

ml التحليل العاملي بطريقة الأرجحية العظمى.

التحليل العاملي للمكونات يستخرج المكونات الرئيسة من مصفوفة ارتباط معدلة، والتي فيها القطر الرئيس يتألف من تقديرات قيمة شيوع بدلاً من 1، خيارات الأمر **factor** وهي **pf** و**ipf** تقوم بحساب التحليل العاملي الرئيس، وهي تختلف في كيفية تقدير قيم الشيوع:

pf تقديرات قيم الشيوع تساوي R^2 من انحدار كل متغير على كل المتغيرات الأخرى.

ipf التقديرات التكرارية لقيم الشيوع.

وتجدر الإشارة إلى أن تحليل المكونات الرئيسة يركز على شرح تباين المتغيرات، إلا أن التحليل العاملي الرئيس يشرح الارتباط بين المتغيرات، سوف نقوم بتطبيق التحليل العاملي الرئيس مع قيم الشيوع المتكررة **(ipf)** على البيانات الخاصة بالكواكب:

```
.factor rings logdsun - logdense, ipf
```

```
(obs=9)

Factor analysis/correlation                      Number of obs    =        9
    Method: iterated principal factors           Retained factors =        5
    Rotation: (unrotated)                        Number of params =       15

    Beware: solution is a Heywood case
          (i.e., invalid or boundary values of uniqueness)
```

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	4.59663	3.46817	0.7903	0.7903
Factor2	1.12846	1.05107	0.1940	0.9843
Factor3	0.07739	0.06438	0.0133	0.9976
Factor4	0.01301	0.01176	0.0022	0.9998
Factor5	0.00125	0.00137	0.0002	1.0000
Factor6	-0.00012	.	-0.0000	1.0000

```
LR test: independent vs. saturated:  chi2(15) =  100.49 Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances
```

Variable	Factor1	Factor2	Factor3	Factor4	Factor5	Uniqueness
rings	0.9760	0.0665	0.1137	-0.0206	-0.0223	0.0292
logdsun	0.6571	-0.6705	0.1411	0.0447	0.0082	0.0966
lograd	0.9267	0.3700	-0.0450	0.0486	0.0166	-0.0004
logmoons	0.9674	-0.0107	0.0078	-0.0859	0.0160	0.0564
logmass	0.8378	0.5458	0.0056	0.0282	-0.0071	-0.0007
logdense	-0.8460	0.4894	0.2059	-0.0061	0.0100	0.0022

Variable	Uniqueness
rings	0.0292
logdsun	0.0966
lograd	-0.0004
logmoons	0.0564
logmass	-0.0007
logdense	0.0022

في الجدول أعلاه، برنامج ستاتا يعرض تحذير "تحذير: الحل عبارة عن حالة هيوود" "Beware: solution is a Heywood case." وعند النقر على الرابط Heywood case في التحذير سوف يتم عرض شرح للمشكلة الموجودة في هذا المثال، وهذا يعكس صغر حجم العينة الاستثنائي ($n = 9$)، وللتبسيط سوف

نستمر في التحليل، ولكن عند إجراء أي بحث، فإن مثل هـذا النـوع مـن التحذيرات يجب أن يعتبر تنبيهاً لنا لإعادة النظر في الطريقة التي نستخدمها.

نرى أن هناك عاملين فقط لهما قيم كامنـة أكثـر مـن 1، باسـتخدام الخيارات **pcf** و**pf** يمكننا تجاهل العوامل البسيطة، وعند استخدام الخيار **ipf** فإنه يجب علينا تحديد كم عدد العوامل التي نريد حفظها، ثم نكرر التحليل مع هذا العدد من العوامل، الآن سوف نقوم بحفظ عاملين اثنين:

```
.factor rings logdsun - logdense, ipf factor(2)
```

```
(obs=9)

Factor analysis/correlation                    Number of obs    =      9
    Method: iterated principal factors         Retained factors =      2
    Rotation: (unrotated)                      Number of params =     11

    Beware: solution is a Heywood case
           (i.e., invalid or boundary values of uniqueness)
```

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	4.57495	3.47412	0.8061	0.8061
Factor2	1.10083	1.07631	0.1940	1.0000
Factor3	0.02452	0.02013	0.0043	1.0043
Factor4	0.00439	0.00795	0.0008	1.0051
Factor5	-0.00356	0.02182	-0.0006	1.0045
Factor6	-0.02537	.	-0.0045	1.0000

```
LR test: independent vs. saturated:  chi2(15) =  100.49 Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances
```

Variable	Factor1	Factor2	Uniqueness
rings	0.9747	0.0537	0.0470
logdsun	0.6533	-0.6731	0.1202
lograd	0.9282	0.3605	0.0086
logmoons	0.9685	-0.0228	0.0614
logmass	0.8430	0.5462	-0.0089
logdense	-0.8294	0.4649	0.0960

بعد إجراء تحليل العوامل مع الخيار **ipf**، يمكننا إنشاء متغيرات مركبة باستخدام الأمر **rotate** والأمر **predict** بنفس الطريقة التي قمنا بها سابقاً، وبسبب مشكلة حالة هيوود Heywood-case فإن القيم العاملية هنا أقل قبولاً من نتائج **pcf** التي تحصلنا عليها سابقاً، وكاستراتيجية للبحث، فإنه من المفيد تكرار التحليل العاملي باستخدام طرق مختلفة حتى نحصل على نتائج أكثر قبولاً.

التحليل العاملي بطريقة الأرجحية العظمى :

Maximum-Likelihood Factoring

التحليل العاملي بطريقة الأرجحية العظمى – يختلف عن خيارات الأمر **factor** – يعتبر وسيلة لإجراء اختبارات لفرضيات، وهذه الاختبارات تساعد في تحديد عدد مناسب من العوامل، وللحصول على عامل أرجحية عظمى واحد لبيانات الكواكب نقوم بطباعة الأمر التالي:

```
.factor rings logdsun - logdense, ml nolog
 factor(1)
(obs=9)

Factor analysis/correlation                    Number of obs    =        9
    Method: maximum likelihood                 Retained factors =        1
    Rotation: (unrotated)                      Number of params =        6
                                               Schwarz's BIC    =  97.8244
    Log likelihood = -42.32054                 (Akaike's) AIC   =  96.6411
```

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	4.47258	.	1.0000	1.0000

```
LR test: independent vs. saturated:  chi2(15) =  100.49 Prob>chi2 = 0.0000
LR test:     1 factor vs. saturated:  chi2(9)  =   51.73 Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances
```

Variable	Factor1	Uniqueness
rings	0.9873	0.0254
logdsun	0.5922	0.6493
lograd	0.9365	0.1229
logmoons	0.9589	0.0805
logmass	0.8692	0.2445
logdense	-0.7715	0.4049

مخرجات الخيار **ml** تحتوي على اختبارين χ^2 لمعدل الاحتمال:

اختبار LR: الاستقلالية مقابل التشبع

وهو يختبر ما إذا كان نموذج ما بدون عامل (مستقل) يتناسب مع مصفوفة الارتباط المشاهدة بدرجة أسوأ بكثير من نموذج متشبع أو نموذج متناسب بشكل كامل، الاحتمال المنخفض (هنا 0.0000 وهذا يعني $p< 0.00005$) يشير إلى أن نموذجاً بدون عامل سوف يكون بسيطاً جداً.

اختبار LR: معامل 1 مقابل التشبع

وهو يختبر ما إذا كان نموذج العامل الواحد الحالي يتناسب بدرجة أسوأ بكثير من نموذج متشبع، قيمة p المنخفضة هنا تشير إلى عامل واحد سوف يكون بسيطاً جداً.

بالطبع، فإن نموذج عاملين سوف يكون أفضل:

```
.factor rings logdsun - logdense, ml nolog
 factor(2)
```

```
(obs=9)

Factor analysis/correlation                    Number of obs    =         9
    Method: maximum likelihood                 Retained factors =         2
    Rotation: (unrotated)                      Number of params =        11
                                               Schwarz's BIC    =   36.6881
    Log likelihood = -6.259338                 (Akaike's) AIC   =   34.5187

    Beware: solution is a Heywood case
            (i.e., invalid or boundary values of uniqueness)
```

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	3.64200	1.67115	0.6489	0.6489
Factor2	1.97085	.	0.3511	1.0000

```
LR test: independent vs. saturated:  chi2(15) =  100.49 Prob>chi2 = 0.0000
LR test:   2 factors vs. saturated:  chi2(4)  =    6.72 Prob>chi2 = 0.1513
(tests formally not valid because a Heywood case was encountered)
```

Factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Uniqueness
rings	0.8655	-0.4154	0.0783
logdsun	0.2092	-0.8559	0.2236
lograd	0.9844	-0.1753	0.0003
logmoons	0.8156	-0.4998	0.0850
logmass	0.9997	0.0264	0.0000
logdense	-0.4643	0.8857	0.0000

الآن نحن وجدنا ما يلي:

اختبار LR: الاستقلال مقابل التشبع

أول اختبار لم يتغير، حيث إنه نموذج بدون عامل بسيط جداً.

اختبار LR: عاملان مقابل التشبع

نموذج عاملين ليس سيئاً بدرجة كبيرة ($p = 0.1513$) من نموذج يتناسب بشكل كامل.

هذه الاختبارات تشير إلى أن عاملين اثنين يسمحان بإنشاء نموذج مناسب.

الطرق المتبعة في حساب التحليل العاملي بطريقة الأرجحية العظمى في العادة تؤدي إلى إيجاد حلول لمشكلة هيود أو إظهار نتائج غير واقعية مثل تباين سالب أو تفردية تساوي 0، وعند حدوث ذلك (كما في نموذج عاملين ml في المثال أعلاه) فإن اختبارات χ^2 تفتقر إلى مبررات واضحة، وعند مشاهدة وصف لهذه الاختبارات فإنه بالإمكان تحديد دليل واضح بخصوص العدد المناسب للعوامل.

التحليل العنقودي – 1 : 1 – Cluster Analysis

التحليل العنقودي يشمل عددًا كبيرًا من الطرق التي تصنف المشاهدات في مجموعات أو عناقيد بناءً على الاختلافات في أعداد متغيراتها، وفي أغلب الأحيان فإن التحليل العنقودي يستخدم مدخلاً استكشافياً لتطوير تصنيف عملي بدلاً من اختبار فرضيات محددة مسبقاً. وفي الحقيقة ليست هناك نظرية واضحة تحدد اختبار الفرضيات في طرق التحليل العنقودي الشائعة، أما عدد الخيارات المتوافر في كل خطوة في التحليل يمكن وصفها بأنها مخيفة، لأن هذه الخطوات من الممكن أن تقودنا إلى نتائج مختلفة، هذا الجزء يُعتبر نقطة بداية عن التحليل العنقودي. وسوف نتناول بعض الأفكار الأساسية ونشرحها باستخدام بعض الأمثلة البسيطة، دليل المستخدم *Multivariate Statistics Reference Manual* يوضح الخيارات الكاملة المتوافرة

في هذا الصدد، كما أن دراسة .Everitt *et a* (2001) تغطي هـذه المواضـيع بتفاصيل أكثر، وتتضمن مقارنات بين العديد من طرق التحليل العنقودي.

كل طرق التحليل العنقودي تبدأ ببعض التعريفات عن أوجـه اختلافهـا وتشابهها، فمقاييس الاختلافات تعكس المسافة بين مـشاهدتين أو مجموعـة محددة من المتغيرات. وبصفة عامة مثل هذه المقاييس يتم تصميمها حتى نجد بأن الاختلاف بين مشاهدتين متماثلتين يساوي صـفرًا وأقـصى اخـتلاف للمشاهدات المختلفة يساوي 1، أما مقاييس التشابه فهي تعكس هذا القيـاس، ولذا فإن المشاهدات المتماثلة لها تشابه يساوي 1؛ خيـارات الأمـر **cluster** ببرنامج ستاتا تعطينا عددًا من الخيارات لقياس التشابه والاختلاف. ولتسهيل عملية حساب ذلك، فإن برنامج ستاتا يقوم داخلياً بتحويل التشابه إلى اختلاف يتم حسابه كما يلي:

الاختلاف = 1 − التشابه

مقياس الاختلاف الافتراضي ببرنامج ستاتا للربط المتوسـط، والـربط الشامل، والربط المنفرد، والربط المتردد هو مسافة إقليدس التي يقوم بحسابها الخيار **measure(L2)** فهذا الخيار يحسب المسافة بين المشاهدة i والمشاهدة j كما يلي:

$$\{\sum_k (x_{ki} - x_{kj})^2\}^{1/2}$$

حيث إن x_{ki} تمثل قيمة المتغير x_k للمشاهدة i، x_{kj} قيمة المتغير x_k للمشاهدة j ، والمجموع يكون لكل متغيرات x التي تم إدخالهـا، الخيـارات الأخـرى المتوافرة لقياس الاختلاف (dis) بين المشاهدات التي تعتمد علـى المتغيـرات المستمرة تتضمن مسافة إقليـدس التربيعيـة the squared Euclidean distance (**L2squared** هو الأمر الافتراضي للارتباط المركزي، والارتبـاط الوسـيط وارتباط الأجزاء) ومسافة القيمة المطلقة (**L1**) ومسافة أعلى قيمة (**Linfinity**) وقياس تشابه معامل الارتباط (**correlation**)، أما الخيارات المتوافرة للمتغيرات الثنائية تتضمن مواءمة بسيطة (**matching**) ومُعامِل التشابه الثنائي لجاكـارد Jaccard(**Jaccard**) ومُعاملات أخر ى، أما الخيار **gower** يعمل مـع متغيـرات

ثنائية ومستمرة مختلفة. وللحصول على قائمة كاملة، وشرح مفصل عن خيارات قياس الاختلافات قم بطباعة الأمر **help measure option**.

طرق التحليل العنقودي يمكن تصنيفها تحت فئتين هما: جزئي وهرمي، الطرق الجزئية تقوم بتقسيم المشاهدات في أعداد مقسّمة مسبقاً لمجموعات غير متداخلة، ولدينا طريقتان للقيام بذلك:

متوسطات k العنقودية Cluster kmeans: التحليل العنقودي لمتوسطات k حيث يقوم المستخدم بتحديد عدد العناقيد (K) المراد إنشاؤها، ثم يقوم ستاتا بعد ذلك بإيجاد هذه العناقيد من خلال إجراء بديل، ويقوم بتقييم المشاهدات التي تقع إلى أقرب متوسط.

وسيط k العنقودي Cluster kmedians: التحليل العنقودي لقيم الوسيط k، وهذا التحليل يشبه التحليل العنقودي لمتوسطات k، ولكن هذا التحليل يتم باستخدام الوسيط.

طرق التجزئة يبدو أنها أبسط وأسرع حسابياً من الطرق الهرمية، وضرورة تحديد عدد العناقيد بالضبط مقدماً يُعتبر أحد عيوب العمل الاستكشافي.

الطرق الهرمية تتضمن عملية تحويل المجموعات الصغيرة إلى مجموعات أكبر بطريقة تدريجية. برنامج ستاتا يستخدم طريقة التكتل *agglomerative* في التحليل العنقودي الهرمي وهذه الطريقة تبدأ مع كل مشاهدة ويتم اعتبار كل مشاهدة كمجموعة منفصلة، أقرب مجموعتين يتم دمجهما وتستمر هذه العملية حتى التوقف في نقطة معينة أو حتى يتم وضع كل المشاهدات في مجموعة واحدة، ويتم عرض نتائج التحليل العنقودي الهرمي في شكل بياني شجري يسمى *dendrogram* أو شكل شجري، هناك عدة خيارات متاحة لطرق الربط، والتي تحدد ما الذي يجب أن تتم مقارنته بين المجموعات التي تحتوي على أكثر من مشاهدة:

الربط الفردي العنقودي cluster singlelinkage: التحليل العنقودي للرابط المفرد، يقوم بحساب الاختلاف بين زوج من المشاهدات الأقل اختلافاً في

مجموعتين، وبالرغم من سهولة هذه الطريقة، فـإن مقاومتهـا أقـل للقـيم المتطرفة أو أخطاء القياس. والمشاهدات تميل للانضمام للتحليل العنقـودي مرة واحدة، وهذا يؤدي إلى إنشاء مجموعات طويلة أو غير متوازنة، وتكون مكونات هذه المجموعات غير متجانسة، ولكنها تُربط مع بعـضها بواسـطة مشاهدات وسطية، وهذه المشكلة تسمى *التسلسل*.

الربط المتوسط العنقودي cluster averagelinkage: التحليل العنقـودي للرابط المتوسط، يستخدم هذا التحليل لتحليل الاختلاف المتوسط للمـشاهدات في مجموعتين مؤدياً إلى ظهور خصائص متوسـطة بـين الـربط الكلـي complete linkage والربط المفرد single linkage، دراسات المحاكاة وجـدت بأن الربط المتوسط يعمل بشكل جيد في العديد من الحالات، وهـذا الـربط موثوق بدرجة معقولة (انظر دراسـة Everitt *et al.* 2001 والمـصادر التـي ذكرتها الدراسة في المراجع)، وهذا النوع من الربط يُستخدم بشكل كبير في علم الآثار.

الربط الشامل العنقودي cluster completelinkage: التحليـل العنقـودي للربط الشامل، وهذا التحليل يستخدم زوج المشاهدات الأقـل تـشابهاً فـي مجموعتين، وهذا الربط أقل حساسية للقيم المتطرفة من الربط المفرد، ولكنه في الاتجاه المعاكس نحو تجميع العديد من المشاهدات في عناقيـد مدمجـة مكانياً ومُحكمة.

الربط المتردد العنقودي cluster waveragelinkage: التحليل العنقـودي للربط المترند الموزون.

الربط الوسيط العنقودي cluster medianlinkage: التحليـل العنقـودي لربط الوسيط.

الربط المتوسط الموزون، والربط الوسيط هـي تباينـات فـي الـربط المتوسط والربط المركزي على التوالي، وفي الحالتين فإن الاخـتلاف فـي كيفية التعامل مع مجموعتين مختلفتين في الحجم عند دمجهما، ففـي الـربط المركزي والربط المتوسط عدد العناصر لكل مجموعة تؤخذ فـي الاعتبـار

عند الاحتساب معطياً في المقابل التأثير الأكبر للمجموعة الأكبر (لأن كل مشاهدة لها نفس الوزن)، وفي الربط الوسيط والربط المتوسط الموزون، فإن المجموعتين يتم إعطاؤهما أوزاناً متساوية بغض النظر عن كيفية وجود العديد من المشاهدات في كل مجموعة. الربط الوسيط يشبه الربط المركزي، حيث إن كلا الرابطين لهما فترات ثقة.

الربط المركزي العنقودي cluster centroidlinkage: التحليل العنقودي للربط المركزي، وهذا النوع من التحليل يدمج المجموعات ذات المتوسطات المتقاربة (بعكس الربط المتوسط والذي يعتني بالمسافة المتوسطة بين عناصر مجموعتين)، هذه الطريقة لها نقاط فترات ثقة حيث الدمج يبدأ عند مستوى أقل للاختلاف من الدمج السابق، الانعكاسات تشير إلى عدم استقرار تركيبة التحليل العنقودي، وصعوبة التفسير، ولايمكن تمثيله بيانياً بواسطة الأمر **cluster dendrogram**.

ربط الأجنحة العنقودية cluster wardslinkage: التحليل العنقودي لرابط الأجنحة، حيث يقوم بتجميع مجموعتين معاً تؤديان إلى أقل زيادة في خطأ مجموع المربعات، ويتناسب هذا التحليل بشكل كبير مع المجموعات التي لها توزيع طبيعي متعدد المتغيرات، وذات أحجام متشابهة، ولا يتناسب هذا التحليل مع المجموعات التي تكون أعداد مشاهداتها في العناقيد أو المجموعات غير متساوية.

سابقاً في هذا الفصل، رأينا أن التحليل العاملي للمكونات الرئيسة للبيانات الموجودة بالملف *planets.dta* (الشكل 3.11) قام بتحديد ثلاثة أنواع من الكواكب: كواكب صخرية داخلية، كواكب عملاقة غازية، وكوكب بلوتو في فئة خاصة به وحده؛ التحليل العنقودي يعتبر طريقة بديلة للسؤال عن نوع الكوكب، وحيث إن المتغيرات مثل عدد الأقمار (*moons*) وتكتلها في الكيلوجرامات (*mass*) يتم قياسه بوحدات قياس غير متشابهة مع اختلاف كبير جداً في التباين، لذلك يجب علينا إجراء نوع من القياس المعياري بطريقة ما لتفادي الحصول على نتائج متأثرة بالتباين الكبير لبعض العناصر، الخيار الأكثر شيوعاً للقيام بذلك – بالرغم من أنه ليس خيارًا تلقائياً – هو وضع متوسط معياري يساوي صفرًا واستخدام وحدة

للانحراف المعياري، يمكن القيام بذلك عن طريق الأمر **egen** (واستخدام متغيرات في شكل لوغاريتمي لنفس الأسباب التي تم مناقشتها سابقاً)، الأمر **summarize** يؤكد أن المتغيرات الجديدة *z* لها متوسطات تساوي صفرًا تقريباً، وانحرافات معيارية تساوي واحد.

```
.egen zrings = std(rings)
.egen zlogdsun = std(logdsun)
.egen zlograd = std(lograd)
.egen zlogmoon = std(logmoons)
.egen zlogmass = std(logmass)
.egen zlogdens = std(logdense)
.summ z*
```

Variable	Obs	Mean	Std. Dev.	Min	Max
zrings	9	-1.99e-08	1	-.8432741	1.054093
zlogdsun	9	-1.16e-08	1	-1.393821	1.288216
zlograd	9	-3.31e-09	1	-1.3471	1.372751
zlogmoon	9	0	1	-1.207296	1.175849
zlogmass	9	-4.14e-09	1	-1.74466	1.365167
zlogdens	9	-1.32e-08	1	-1.453143	1.128901

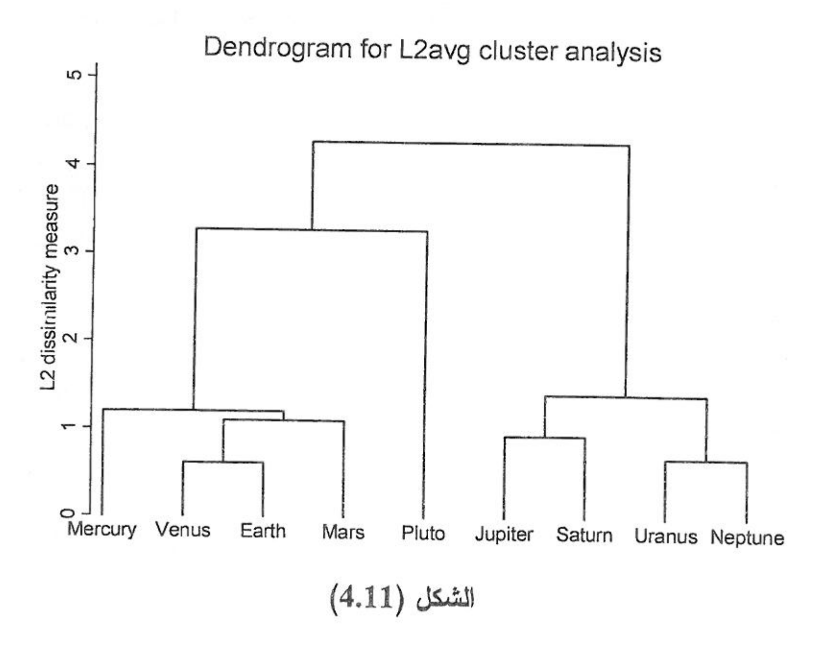
نتيجة التحليل العاملي للمكونات الرئيسة للأنواع الثلاثة هي نتيجة موثوقة، كما يمكن الحصول على نفس النتيجة من خلال التحليل العنقودي. فعلى سبيل المثال، يمكننا إجراء تحليل عنقودي هرمي مع متوسط الرابط باستخدام مسافة إقليدس Euclidean distance(L2) كمقياس للاختلاف. الخيار **name(*L2age*)** يعطي اسماً للنتائج التي نحصل عليها من هذا التحليل حتى يمكننا الإشارة إلى هذه النتائج في الأوامر اللاحقة، خاصية إعطاء أسماء للنتائج هي شيء مفيد عندما نحتاج إلى إجراء عدد من عمليات التحليل العنقودي ومقارنتها مع المخرجات.

```
.cluster averagelinkage zrings zlogdsun zlograd
 zlogmoon zlogmass
        zlogdens, measure(L2) name(L2avg)
```

لم يحدث أي شيء بالرغم من أننا قد لاحظنا أن البيانات الآن تحتوي على ثلاثة متغيرات جديدة مع أسماء تم إنشاؤها بناءً على المتغير *L2avg*، هذه المتغيرات الجديدة *L2avg** ليست مهمة بشكل مباشر ولكن يمكن

استخدامها بشكل غير مباشر بواسطة الأمــر **cluster dendrogram** لرســم شجرة التحليل العنقودي أو الشكل الشجري الذي يعرض نتائج أحدث تحليــل عنقودي هرمي (الشكل 4.11)، الخيار **label(*planet*)** يؤدي إلى إظهار أسماء الكواكب (قيم المتغير *planet*) كتوصيفات أسفل الرسم البياني.

```
.cluster dendrogram, label(planet) ylabel(0(1)5)
```



الشكل (4.11)

الشكل الشجري (4.11) يعتبر أداة تفسيرية رئيسة للتحليــل العنقــودي الهرمي، ويمكننا تتبع العملية المتجمعة من كل مــشاهدة ومقارنــة نتيجتهــا العنقودية المتميزة – التي تظهر في أسفل الشكل – مع كل النتائج المدمجــة في عنقود واحد في الأعلى، كوكب فينوس Venus وكوكب الأرض Earth من ناحية، وكذلك كوكبا أورانوس Uranus ونيبتون Neptune من ناحية أخــرى، الأقل اختلافاً أو أكثر الأزواج تشابهاً، حيث تم تجميعهــا أولاً وكونــوا أول عنقودين من مشاهدات متعددة في مستوى (الاختلاف) أقل مــن 1، كوكبــا جوبتير Jupiter وزحل Saturn ثم الأرض Earth – فينوس Venus والمــريخ Mars ثم فينوس Venus – الأرض Earth – المريخ Mars وعطارد Mercury

وأخيراً المشتري Jupiter – زحـــل Saturn وأورانـــوس Uranus – نيبتـــون Neptune تم تجميعها في سلسلة سريعة وجميعها لها اختلافات أكثـــر مـــن 1 بقليل، عند هذه النقطة لدينا نفس الثلاث مجموعات التـــي أظهرهـــا تحليـــل المكونات الرئيسة في الشكل (3.11): الكواكب الصخرية الداخلية والكواكب العملاقة الغازية وبلوتو، الثلاثة عناقيد تبقى مستقرة حتـــى الوصـــول إلـــى مستوى أعلى من الاختلاف (أعلى من 3) بلوتو تم دمجه ضـــمن مجموعـــة الكواكب الصخرية الداخلية، وعند الوصول إلى درجة اختلاف أعلى مـــن 4 فسوف يندمج آخر عنقودين.

إذن كم نوعاً من الكواكب لدينا؟ الشكل (4.11) يعطي الإجابة بوضوح وهي "الإجابة التي تعتمد على" كم درجة الاختلاف التي نريد القبول بها لكل نوع؟ الخطوط العمودية الطويلة بين مرحلة العناقيد الثلاثة ومرحلة العنقودين في أعلى جزء من الرسم البياني تشير إلى وجود ثلاثة أنواع يمكن التمييـــز بينها بوضوح، ويمكننا إنقاص هذه الأنواع الثلاثة إلى نوعين فقط وذلك بدمج مشاهدة (بلوتو Pluto) لأنه مختلف عن البقية في مجموعته، ويمكننا التوســـع ليكون لدينا خمسة أنواع فقط، وذلك من خلال توضيح الفرق بين عدد مـــن الكواكب (على سبيل المثال، عطـــارد وفينـــوس Venus – الأرض Earth – المريخ Mars) والتي لاتختلف كثيراً بحسب معايير المجموعة الشمسية، ولذا فإن الشكل الشجري يدعم الشكل الذي يحتوي على ثلاثة أنواع.

الأمر **cluster generate** يقوم بإنشاء متغير جديد يشير إلـــى النـــوع أو المجموعة التي تنتمي إليها كل مشاهدة. في هذا المثـــال **group(3)** يتطلـــب ثلاث مجموعات، الخيار **name(*L2avg*)** يقوم بتحديد نتـــائج معينـــة ونقـــوم بإعطائها اسم *L2avg*. هذا الخيار هو أكثر الخيارات فائدة عنـــدما تتـــضمن الحسابات تحليلاً عنقوياً متعددًا.

```
.cluster generate plantype= groups(3),
 name(L2avg)
.label variable plantype "Planet type"
.list planet plantype
```

	planet	plantype
1.	Mercury	1
2.	Venus	1
3.	Earth	1
4.	Mars	1
5.	Jupiter	3
6.	Saturn	3
7.	Uranus	3
8.	Neptune	3
9.	Pluto	2

الكواكب الصخرية الداخلية تم ترميزها لتكون 1 = *plantype*، والكواكب العملاقة الغازية 3 = *plantype*، وكوكب بلوتو وحده 2 = *plantype*، ترميز المجموعات تكون 3، 2، 1 من اليسار إلى اليمين مرتباً بحيث تكون العناقيد النهائية في الشكل البياني الشجري (الشكل 4.11)، وعند القيام بحفظ البيانات فإن الرموز يمكن استخدامها كمتغير تصنيفي آخر في أي تحليلات لاحقة.

بيانات الكواكب لها نمط طبيعي قوي وهذا هو السبب وراء حصولنا على نتائج متشابهة من استخدام تقنيات تحليل مختلفة مثل تحليل المكونات الرئيسة، والتحليل العنقودي. يمكننا اختيار طرق أخرى لقياس الاختلاف وطرق ربط أخرى لهذا المثال، وسوف نصل إلى نفس النتيجة تقريباً. ومن ناحية أخرى، فإن البيانات النمطية الضعيفة أو المعقدة في العادة تُظهِر نتائج مختلفة بناءً على طريقة التحليل المستخدمة. العناقيد التي تَنتُج من طريقة تحليل واحدة قد لا تكون دليلاً على إمكانية تكرار نفس النتائج مع طرق أخرى أو مع قرارات تحليلية مختلفة قليلاً.

التحليل العنقودي – 2 : 2 – Cluster Analysis

اكتشاف تصنيف موثوق وبسيط لتوصيف الكواكب التسعة كان واضحاً، ولعرض أمثلة أكثر صعوبة، فإننا سوف نستخدام بيانات الدول الموجودة بالملف *Nations2.dta*، متغيرات التنمية البشرية للأمم المتحدة يمكن تطويرها لتكون تصنيفاً عملياً للدول.

```
.use C:\data\Nations2.dta, clear
```

.describe

```
Contains data from C:\data\Nations2.dta
  obs:          194                          UN Human Development Indicators
 vars:           13                          2 Jul 2012 06:11
 size:       12,804
```

variable name	storage type	display format	value label	variable label
country	str21	%21s		Country
region	byte	%8.0g	region	Region
gdp	float	%9.0g		Gross domestic product per cap 2005$, 2006/2009
school	float	%9.0g		Mean years schooling (adults) 2005/2010
adfert	float	%8.0g		Adolescent fertility: births/1000 fem 15-19, 2010
chldmort	float	%9.0g		Prob dying before age 5/1000 live births 2005/2009
life	float	%9.0g		Life expectancy at birth 2005/2010
pop	float	%9.0g		Population 2005/2010
urban	float	%9.0g		Percent population urban 2005/2010
femlab	float	%9.0g		Female/male ratio in labor force 2005/2009
literacy	float	%9.0g		Adult literacy rate 2005/2009
co2	float	%9.0g		Tons of CO2 emitted per cap 2005/2006
gini	float	%9.0g		Gini coef income inequality 2005/2009

Sorted by: region country

بالعمل مع نفس البيانات في الفصل (7) رأينا تحويلات غير خطيّة مثل اللوغاريتمات التي ساعدت في جعل التوزيعات أكثر طبيعية وجعل العلاقات خطيّة أكثر بين بعض المتغيرات، نفس الطرق للتحويلات غير الخطية يمكن تطبيقها في التحليل العنقودي، ولكن لجعل المثال أكثر بساطة فلن نقوم باستخدام هذه الطرق هنا. التحويلات الخطية التي تجعل المتغيرات أكثر معيارية مازالت ضرورية هنا، وإذا لم نقم بذلك فإن المتغير *gdp* والذي يتراوح من حوالي 280 دولارًا إلى 74,906 دولارات (بانحراف معياري 13,942 دولارًا) فإن ذلك قد يؤدي إلى التغلب على المتغيرات الأخرى مثل *life* والذي يتراوح بين 46 تقريباً إلى 83 سنة (بانحراف معياري 10 سنوات). في الجزء السابق قمنا بتحويل بيانات الكواكب لتكون بيانات معيارية من خلال طرح متوسط كل متغير ثم قسمته على انحرافه المعياري حتى تكون كل نتائج *z* لها انحرافات معيارية تساوي 1. في هذا الجزء سوف

نقوم باتباع طريقة مختلفة تسمى مدى المعايرة، والتي تعمل بشكل جيد مع التحليل العنقودي.

مدى المعايرة يتضمن قسمة كل متغير على مداه، ليس هناك أمر ببرنامج ستاتا يقوم بذلك بطريقة مباشرة ولكن يمكننا إنشاء أمر للقيام بذلك، وللقيام بذلك سوف نستخدم النتائج التي قام برنامج ستاتا بحفظها في ذاكرته العاملة بعد الأمر **summarize** وذلك بطباعة الأمر **return list**، (بعد عمليات أوامر النماذج مثل **regress** أو **factor** وبدلاً من ذلك قم باستخدام الأمر **ereturn list**). في هذا المثال سوف نلقي نظرة على النتائج المخزنة بعد الأمر *pop* **summarize** ثم استخدام أعلى وأقل قيم (المخزنة كأعداد قياسية والتي يسميها برنامج ستاتا r(max) وr(min)) لحساب نسخة جديدة لمدى المعايرة للمجتمع.

```
.summarize gdp
```

Variable	Obs	Mean	Std. Dev.	Min	Max
gdp	179	12118.74	13942.34	279.8	74906

```
.return list
scalars:
                  r(N) =  179
              r(sum_w) =  179
               r(mean) =  12118.73919336756
                r(Var) =  194388878.6050418
                 r(sd) =  13942.34121677711
                r(min) =  279.7999877929688
                r(max) =  74906
                r(sum) =  2169254.315612793

.generate rgdp = gdp/(r(max) - r(min))
.label variable rgdp "Range-standardized GDP"
```

يمكن استخدام أوامر مشابهة أخرى لإنشاء نسخ من مدى المعايرة لمتغيرات ظروف المعيشة:

```
.quietly summ school
.generate rschool = school/(r(max) - r(min))
.label variable rschool "Range-standardized schooling"
```

```
.quietly summ adfert
.generate radfert = adfert/(r(max) - r(min))
.label variable radfert "Range-standardized
adolescent fertility"
```

وهكذا معرّفاً المتغيرات 8 الجديدة والتي تظهر بالقائمة أدناه.

```
.describe rgdp-rfemlab
```

variable name	storage type	display format	value label	variable label
rgdp	float	%9.0g		Range-standardized GDP
rschool	float	%9.0g		Range-standardized schooling
radfert	float	%9.0g		Range-standardized adolescent fertility
rfemlab	float	%9.0g		Range-standardized female labor
rchldmort	float	%9.0g		Range-standardized child mortality
rlife	float	%9.0g		Range-standardized life expectancy
rpop	float	%9.0g		Range-standardized population
rurban	float	%9.0g		Range-standardized percent urban

إذا لم تكن أوامر **generate** التي تم استخدامها سابقاً صحيحة، فإن مدى متغيرات مدى المعايرة الجديدة يجب أن يساوي 1، الأمر **tabstat** يؤكد ذلك.

```
.tabstat rgdp - rfemlab, statistics(range)
```

stats	rgdp	rschool	radfert	rfemlab	rchldm~t	rlife	rpop	rurban
range	1	.9999999	1	1	1	1	1	.9999999

عند تحويل المتغيرات التي تهمنا إلى الصيغة المعيارية يمكننا إجـــراء التحليل العنقودي، وعند قيامنا بتصنيف أكثر من 100 دولة إلى عدة أنـــواع فإنه ليس لدينا أي سبب لافتراض أن كل نوع سوف يتـــضمن نفـــس عـــدد الدول، الرابط المتوسط (الذي تم استخدامه مع بيانات الكواكب) مع بعـــض الطرق الأخرى يعطي كل مشاهدة نفس الوزن، هذا يجعـــل العناقيـــد أكبـــر وأكثر تأثيراً أثناء عمليات التجميع. ومن ناحية أخرى، فإن طرق المتوســـط الموزون والرابط المتوسط تعطي وزناً متساوياً لكل عنقود بغض النظر عن عدد المشاهدات التي يحتويها. وبالتالي فإن مثل هذه الطرق تعمل بشكل جيد للكشف عن العناقيد غير متساوية الحجم، الـــرابط الوســـيط يـــشبه الـــرابط المركزي حيث إنه عرضة للانعكاسات (والتي سوف تحدث مع هذه البيانات)

لذلك فإن المثال التالي يقوم بتطبيق الربط المتوسط الموزون، كما أن مسافة القيمة المطلقة ((measure(L1 تعتبر طريقة لقياس الاختلاف.

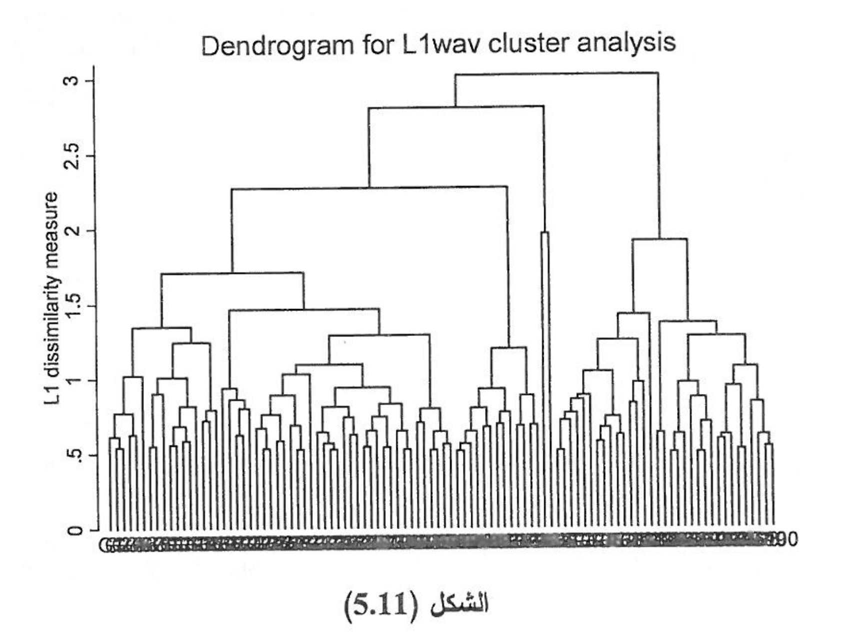
```
.cluster   waveragelinkage   rgdp   -   rfemlab,
 measure(L1) name(L1wav)
```

التحليل العنقودي الكامل يقوم بإنتاج شكل شجري كبير جداً:

```
.cluster dendrogram
too many leaves; consider using the cutvalue()
or cutnumber() options r(198)؛
```

بعد ظهور رسالة الخطأ أعلاه، فإن الشكل (5.11) يقوم باستخدام الخيار cutnumber(100) لإنشاء شكل شجري يبدأ مع 100 مجموعــة بعــد فتــرة بسيطة من بداية أول عمليات الدمج.

```
.cluster dendrogram, ylabel(0(.5)3)
 cutnumber(100)
```



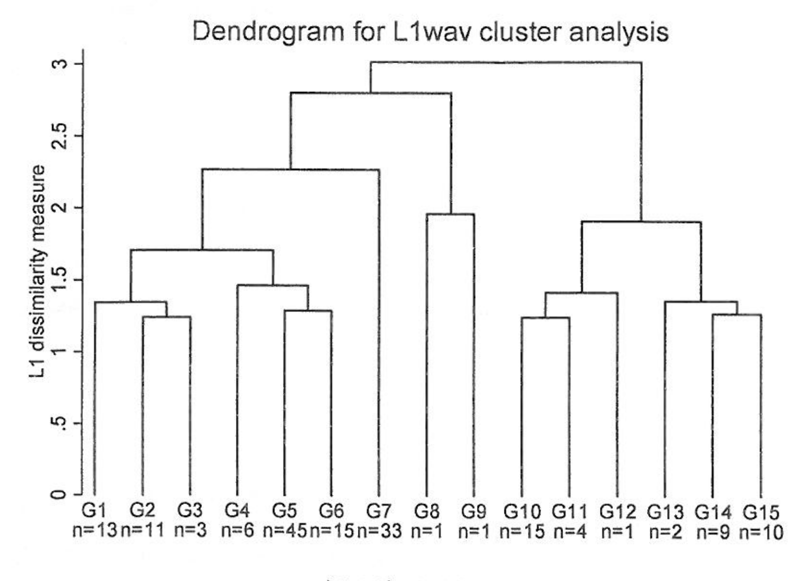
الشكل (5.11)

لا يمكن قراءة التوصيفات الموجودة أسفل الشكل (5.11)، ولكن يمكننــا تتبع الاتجاه العام للتحليل العنقودي. أغلب عمليات الدمج تمت عند اختلافــات أقل من 1. هناك دولتان فقط في مركز الشكل الشجري وهذا غير معتاد. هاتان

الدولتان قامتا بمقاومة الدمج حتى درجة اختلاف تساوي 2 تقريباً، حيث إن هاتين الدولتين المستقرتين تختلفان عن كل الدول الأخرى، هذا واحد من أربعة عناقيد باقية في مستوى الاختلاف الأعلى من 2، الأول والرابع من هذه العناقيد الأربعة (اقرأ من اليسار إلى اليمين) تظهر غير متجانسة وتم إنشاؤها من خلال عمليات دمج متتالية لعدد من المجموعات الفرعية الرئيسة الواضحة. وعلى خلاف ذلك فإن العنقود الثاني يظهر أكثر تجانساً، حيث إنه يجمع بين العديد من الدول التي تم دمجها في مجموعتين فرعيتين لهما اختلاف أقل من 1 ثم دمجهما في مجموعة واحدة عند مستوى اختلاف أعلى من 1 بقليل.

الشكل (6.11) يعرض وضعية أخرى لهذا التحليل، وهذه المرة باستخدام الخيار **cutvalue(1.2)** لعرض العناقيد فقط التي يكون اختلافها أعلى من 1.2 بعد القيام بأغلب عمليات الدمج، الخيار **showcount** يقوم بتحويل التوصيفات أسفل الشكل البياني (n=13, n=11, etc.) وهي تشير إلى عدد الدول في كل مجموعة، حيث إننا نرى أن المجموعات 8، 9، 12 تحتوي كل منها على دولة واحدة فقط.

```
.cluster dendrogram, ylabel(0(.5)3)
 cutvalue(1.2) showcount
```



الشكل (6.11)

الشكل (6.11) يوضح بأن هناك 15 مجموعة باقية عند درجة اختلاف أعلى من 1.2، للتوضيح فقط سوف نأخذ في الاعتبار أعلى أربع مجموعات فقط والتي لها درجة اختلاف أعلى من 2، الأمر **cluster generate** يقوم بإنشاء متغير تصنيفي لكل المجموعات الأربع النهائية من التحليل العنقودي ونقوم بتسميته *L1wav*.

```
.cluster generate ctype = groups(4),
 name(L1wav)
.label variable ctype "Country type"
```

وسوف نقوم في الخطوة التالية باختبار ماهي الدول التي تعود لكل مجموعة وذلك من خلال طباعة الأمر:

```
.sort ctype
.by ctype: list country
```

القائمة الطويلة الناتجة من الأمر أعلاه – لم يتم عرضها هنا – توضح بأن هناك عنقودًا به دولتان لاحظنا وجودهما في الشكل (5.11) من النوع 3 وهما الهند والصين، والعنقود الثاني المتجانس نسبياً في الشكل (5.11) والذي يُصنف كنوع 4 يتضمن مجموعة كبيرة من أفقر الدول وأغلبها دول أفريقية، النوع 2 وهو متنوع نسبياً يحتوي على الدول الأكثر ثراءً، وهي تتضمن الولايات المتحدة واليابان والعديد من الدول الأوروبية، النوع 1 متنوع أيضاً يتضمن الدول التي بها مستويات معيشة متوسطة؛ ويظل السؤال الموضوعي هو ما إذا كانت هذه الأنواع ذات معنى، وهذا السؤال ليس إحصائياً، والإجابة تعتمد على الاستخدامات. وما هي التصنيفات التي نحتاج إليها، الاختيار بين مجموعة مختلفة من الخيارات في خطوات التحليل العنقودي سوف يؤدي إلى الحصول على نتائج مختلفة، وبإجراء التحليل مع مجموعة متعددة من الخيارات سوف نتعرف على أكثر النتائج استقراراً.

استخدام الدرجات العاملية في الانحدار:

Using Factor Scores in Regression

تحليل المكونات الرئيسة، والتحليل العاملي في العادة يساعدان على تعريف المتغيرات المركبة الجديدة لإجراء عمليات تحليل أكثر، فمثلاً

الدرجات العاملية والتي يتم حسابها بالأمر **predict** يمكن أن تصبح متغيرات مستقلة أو غير مستقلة في تحليل الانحدار اللاحق. ولتوضيح ذلك سوف نستخدم بيانات الاستقصاء الموجودة بالملف *PNWsurvey2_11.dta*.

```
.use C:\data\PNWsurvey2_11.dta, clear
.describe
```

```
Contains data from C:\data\PNWsurvey2_11.dta
  obs:           734                          Pacific NW CERA survey (February 2011)
 vars:            16                          2 Jul 2012 06:11
 size:        13,946
```

variable name	storage type	display format	value label	variable label
age	byte	%8.0g	age	Age in years
sex	byte	%8.0g	sex	Gender
educ	byte	%14.0g	degree	Highest degree completed
party	byte	%11.0g	party	Political party identification
newcomer	byte	%9.0g	yesno	Moved here within past 5 years
surveywt	float	%9.0g		CERA survey wt--adults/age/race/sex/county
forest	byte	%13.0g	eff	Loss of forestry jobs or income
cutting	byte	%13.0g	eff	Overharvesting or heavy cutting of timber
losfish	byte	%13.0g	eff	Loss of fishing jobs or income
overfish	byte	%13.0g	eff	Overfishing in the ocean
water1	byte	%13.0g	eff	Water quality or pollution issues
water2	byte	%13.0g	eff	Water supply problems
warming	byte	%13.0g	eff	Global warming or climate change
sprawl	byte	%13.0g	eff	Urban sprawl/development of countryside
weather	byte	%13.0g	eff	Unusual/extreme weather-related events
losscen	byte	%13.0g	eff	Loss of scenic natural beauty

```
Sorted by:
```

المتغيرات 16 في هذه البيانات تمثل بيانات من استقصاء هاتفي للسكان في منطقة ساحلية بشمال غرب المحيط الهادئ. هذا الاستقصاء والذي تم إجراؤه في فبراير 2011 يعتبر جزءًا من مبادرة تتضمن سلسلة استقصاءات تم القيام بها تحت إشراف منظمة المجتمع والبيئة في المناطق الريفية الأمريكية والتي تُعرف اختصاراً باسم (CERA) (انظر دراسة Hamilton et al. 2010b; Safford and Hamilton 2010).

في هذا المثال، هناك عشرة أسئلة تبدأ من *forest* وحتى *losscen* تستفسر عما إذا كانت قضايا بيئية معينة لها تأثيرات محلية. الأسئلة كانت كما يلي:

سوف أقوم بقراءة قائمة من القضايا البيئية التي قد تكون مشاكل في المناطق الريفية، أما بشأن المكان الذي تعيش فيه فأنا أود أن أعرف ما إذا كنت تعتقد أن هذه القضايا ليس لها تأثير أو تأثيرات بسيطة أو تأثيرات جوهرية على أسرتك أو المجتمع الذي تعيش فيه خلال فترة الـ 5 سنوات الماضية؟

خسارة وظائف بالغابات أو دخل من الغابات؟

الإفراط أو التبذير في قطع أخشاب البناء؟

خسارة وظائف صيد الأسماك أو دخل صيد الأسماك؟

الصيد الجائر في المحيط؟

جودة الماء وقضايا التلوث؟

مشاكل في إمدادات المياه؟

الاحتباس الحراري أو التغير المناخي؟

الزحف العمراني أو التطور السريع في المناطق الريفية؟

الأحداث غير العادية أو البالغة المتعلقة بالطقس؟

خسارة الجمال الطبيعي الخلاب؟

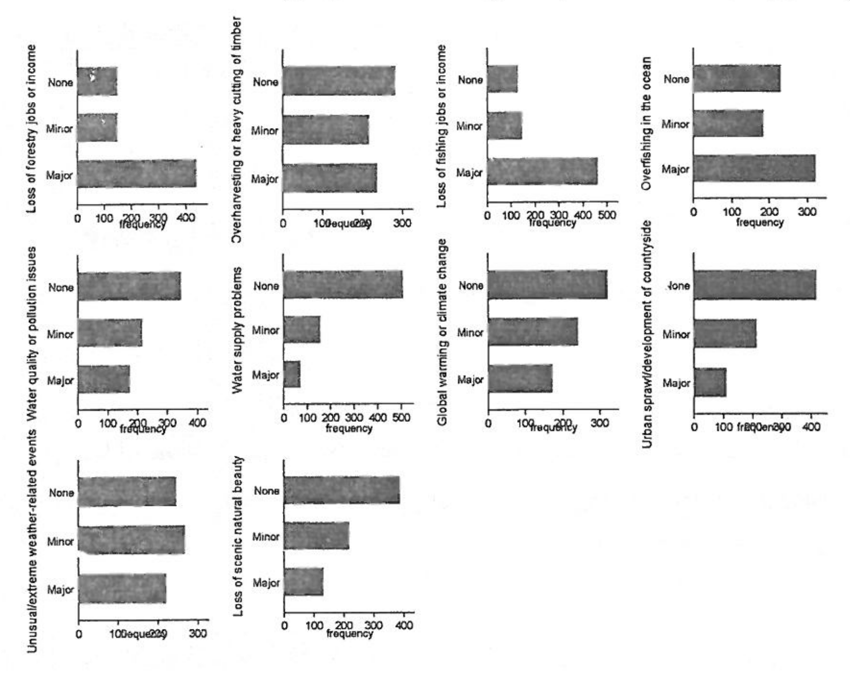
لكل سؤال يمكن للمشارك في الدراسة أن يقول ما إذا كانت هذه القضايا ليس لها تأثير أو تأثيرات بسيطة أو تأثيرات جوهرية على أسرهم أو مجتمعهم.

```
.svy: tab forest
(running tabulate on estimation sample)

Number of strata   =        1          Number of obs      =        734
Number of PSUs     =      734          Population size    = 725.97798
                                       Design df          =        733

-------------------------
Loss of   |
forestry  |
jobs or   |
income    | proportions
----------+--------------
     None |       .2084
    Minor |       .2108
    Major |       .5808
          |
    Total |           1
-------------------------
  Key:  proportions  =  cell proportions
```

العديد من المشاركين أوضحوا بأن الخسارة في وظائف الغابات وصيد الأسماك لها تأثيرات بسيطة، وإمدادات المياه والزحف العمراني وخسارة الجمال الطبيعي الخلاب هو مصدر قلق فوري لمنطقتهم كما يصوّر الشكل (7.11)، هذا الشكل البياني تم إنشاؤه برسم عشرة أشكال لأعمدة بيانية بسيطة باستخدام الأمر **catplot** مع أوزان تحليلية تم إعطاؤها بواسطة المتغير *surveywt* (**[aw =** ***surveywt*****]**، انظر الفصل 4) تم توحيد هذه الأشكال العشرة في صورة واحدة باستخدام الأمر **graph combine**.



الشكل (7.11)

هل القلق حول هذه القضايا البيئية العشر يعكس عددًا أقل من الأبعاد الكامنة؟ التحليل العاملي الأساسي مع قيم الشيوع المتكررة (**ipf**) يُحدد الأبعاد الكامنة التي تأخذ في الاعتبار نمط الارتباط بين المتغيرات.

```
.factor forest cutting losfish overfish water1
 water2 warming sprawl weather losscen, ipf
```

```
(obs=734)

Factor analysis/correlation                        Number of obs    =      734
    Method: iterated principal factors             Retained factors =        9
    Rotation: (unrotated)                          Number of params =       45
```

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	3.34457	2.18016	0.5886	0.5886
Factor2	1.16440	0.78239	0.2049	0.7936
Factor3	0.38201	0.06367	0.0672	0.8608
Factor4	0.31834	0.09420	0.0560	0.9168
Factor5	0.22413	0.06330	0.0394	0.9563
Factor6	0.16083	0.11590	0.0283	0.9846
Factor7	0.04494	0.00940	0.0079	0.9925
Factor8	0.03554	0.02822	0.0063	0.9988
Factor9	0.00732	0.00757	0.0013	1.0000
Factor10	-0.00025	.	-0.0000	1.0000

```
LR test: independent vs. saturated:  chi2(45) = 2030.48 Prob>chi2 = 0.0000
```

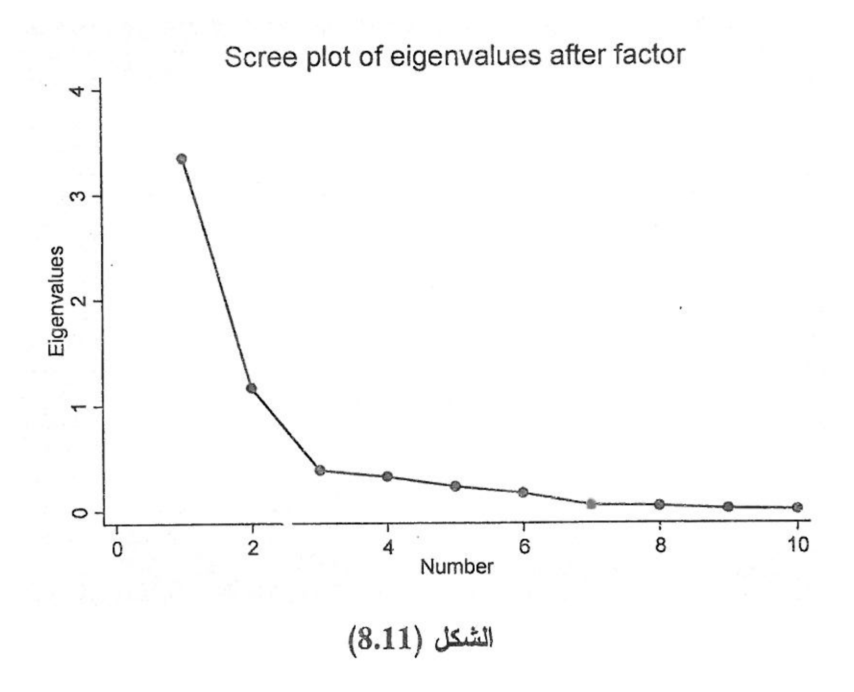
Factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7
forest	0.3408	0.6168	0.0569	0.3357	-0.0288	0.0285	0.0608
cutting	0.6254	0.1541	0.0641	-0.0126	-0.3822	-0.0466	0.0004
losfish	0.4637	0.6746	-0.0195	-0.0834	0.2061	0.0068	-0.0727
overfish	0.6568	0.2032	-0.0637	-0.3887	-0.0030	-0.0670	0.0416
water1	0.6410	-0.1694	-0.2377	0.0438	-0.0795	0.2092	-0.1172
water2	0.6181	-0.2101	-0.3671	0.1051	0.0989	0.0258	0.1055
warming	0.6393	-0.2047	0.2271	-0.0950	0.0374	0.1280	0.0772
sprawl	0.6046	-0.2189	-0.0078	0.1117	0.0683	-0.1921	-0.0294
weather	0.4960	-0.1952	0.3432	0.0485	0.1012	0.1310	-0.0129
losscen	0.6144	-0.2511	0.0976	0.1033	0.0483	-0.1960	-0.0486

Variable	Factor8	Factor9	Uniqueness
forest	0.0347	0.0045	0.3809
cutting	-0.0223	-0.0035	0.4321
losfish	-0.0585	-0.0069	0.2713
overfish	0.0818	0.0049	0.3593
water1	0.0083	0.0041	0.4381
water2	0.0078	-0.0146	0.4061
warming	-0.1150	0.0029	0.4518
sprawl	-0.0163	0.0625	0.5273
weather	0.1005	0.0029	0.5580
losscen	-0.0017	-0.0554	0.4930

كم عدد العوامل التي يجب علينا حفظها؟ أول اثنين فقط قيمة الجذر الكامن لها أعلى من 1 بالرغم من أنها مع التحليل العاملي الأساسي (لا يتشابه مع تحليل المكونات الرئيسة) قطع قيمة الجذر الكامن لها – 1 في بعض الأحيان يُعتبر دقيقة جداً. الرسم البياني للحصى والحجارة (الشكل 8.11) يؤكد بيانياً أنه بعد أول عاملين هناك قيمة جذر كامن منخفضة مع ثبات باقي العوامل الأخرى.

```
.screeplot
```



الشكل (8.11)

بعد تجارب أكثر مع العاملين الاثنين المحتفظ بهما والمدوّرة سوف تكون هناك ثلاثة عوامل أو أكثر (لم يتم عرضها هنا) كما يبدو أن العاملين الاثنين لهما نتائج أفضل وقابلة للتفسير، وهذا شيء مهم يجب أخذه في الاعتبار، عندما نقرر الاستمرار مع العاملين، فإن أول خطوة (لأننا نستخدم التحليل العاملي الرئيس التكراري) تتضمن تكرار التحليل مع الاقتصار على عاملين **.factor(2)**

```
.factor forest cutting losfish overfish water1 water2 warming sprawl weather losscen, ipf factor(2)
(obs=734)

Factor analysis/correlation                    Number of obs    =      734
    Method: iterated principal factors         Retained factors =        2
    Rotation: (unrotated)                      Number of params =       19
```

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	3.20673	2.14023	0.7504	0.7504
Factor2	1.06650	0.85629	0.2496	1.0000
Factor3	0.21021	0.08664	0.0492	1.0492
Factor4	0.12357	0.03148	0.0289	1.0781
Factor5	0.09209	0.06060	0.0215	1.0997
Factor6	0.03149	0.08506	0.0074	1.1070
Factor7	-0.05357	0.05634	-0.0125	1.0945
Factor8	-0.10991	0.00827	-0.0257	1.0688
Factor9	-0.11817	0.05758	-0.0277	1.0411
Factor10	-0.17575	.	-0.0411	1.0000

LR test: independent vs. saturated: chi2(45) = 2030.48 Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Uniqueness
forest	0.3217	0.5026	0.6439
cutting	0.6020	0.1239	0.6223
losfish	0.4788	0.7268	0.2425
overfish	0.6289	0.1824	0.5713
water1	0.6249	-0.1549	0.5855
water2	0.5916	-0.1811	0.6172
warming	0.6305	-0.1924	0.5654
sprawl	0.6074	-0.2205	0.5824
weather	0.4799	-0.1770	0.7384
losscen	0.6155	-0.2515	0.5579

طريقة تدوير المحاور بروماكس Promax (منحـرف) يبسّط الأنماط العاملية، بينما يسمح بدرجة ما من الارتباط بين العوامـل، الارتبـاط بـين العوامل سوف يكون إحصائياً ضعيفاً جداً، لأن هـذه العوامـل لهـا تبـاين متقاطع، وعموماً فقد تكون هذه العوامل أكثر واقعية إذا تم اعتبار أنها تمثل أساساً وليس من الضروري أنها أبعاد لها علاقة بالقضايا البيئية.

```
.rotate, promax

Factor analysis/correlation                        Number of obs    =      734
    Method: iterated principal factors             Retained factors =        2
    Rotation: oblique promax (Kaiser off)          Number of params =       19
```

Factor	Variance	Proportion	Rotated factors are correlated
Factor1	3.06450	0.7171	
Factor2	1.86514	0.4365	

```
LR test: independent vs. saturated:  chi2(45) = 2030.48 Prob>chi2 = 0.0000
```

Rotated factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Uniqueness
forest	-0.0567	0.6164	0.6439
cutting	0.4346	0.2980	0.6223
losfish	-0.0703	0.8951	0.2425
overfish	0.4190	0.3668	0.5713
water1	0.6389	0.0124	0.5855
water2	0.6277	-0.0244	0.6172
warming	0.6686	-0.0253	0.5654
sprawl	0.6674	-0.0612	0.5824
weather	0.5291	-0.0513	0.7384
losscen	0.6949	-0.0915	0.5579

Factor rotation matrix

	Factor1	Factor2
Factor1	0.9662	0.6109
Factor2	-0.2578	0.7917

أسئلة الاستقصاء حول تأثيرات جودة المياه، أو قضايا إمدادات الميـــاه والاحتباس الحراري، والأحداث غير العادية المتعلقــة بـــالطقس والزحـــف العمراني، وخسارة الجمال الطبيعي الخلاب كلها تكون أعلى بكثير في العامل الأول factor 1، والقلق حول خسارة وظائف صيد الأسماك *overfish* ووظائف الغابات *cutting* لها نفس التشبع على العاملين الأول والثاني معاً، وبناءً على هذا النمط يمكننا تفسير العامل الأول factor 1 على أنه يمثل القضايا البيئيـــة

العامة والعامل الثاني factor 2 يمثل القضايا المتعلقة بمصادر الوظائف، الأمر **predict** يقوم بحساب القيم العاملية والتي يتم تعريفها بواسطة المتغيرات المركبة كمجاميع للقيم المعيارية للمتغيرات وأوزانها من خلال قيمها العاملية، المتغيران المركبان الجديدان تم تسميتهما باسم *enviro* و *resjobs*.

```
.predict enviro resjobs

(regression scoring assumed)

Scoring coefficients (method = regression; based on promax(3) rotated factors)
```

Variable	Factor1	Factor2
forest	0.00448	0.15784
cutting	0.12496	0.14577
losfish	0.01218	0.68645
overfish	0.13504	0.09935
water1	0.18799	0.02717
water2	0.16724	0.00984
warming	0.20531	0.00398
sprawl	0.19251	-0.00430
weather	0.11661	0.00008
losscen	0.21054	-0.01245

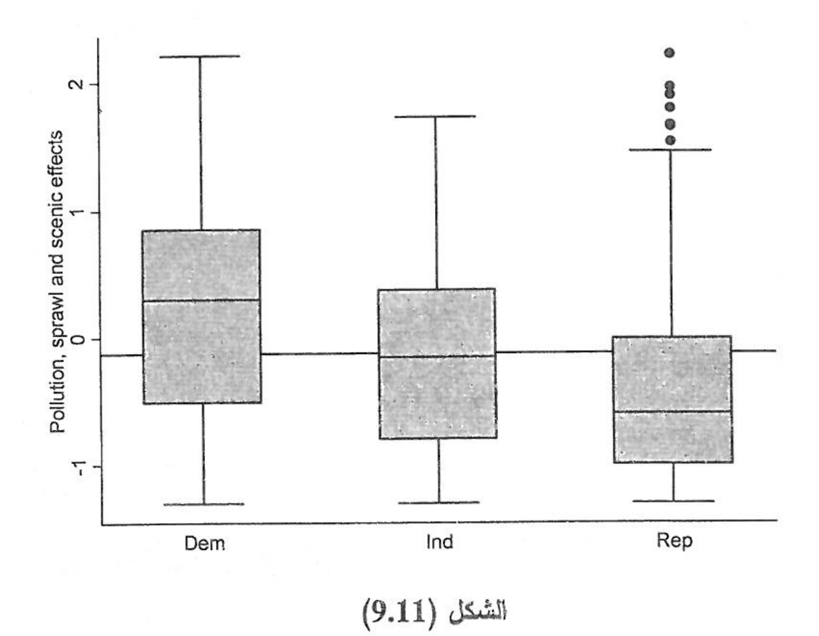
```
.label variable enviro "Pollution, sprawl and scenic effects"
.label variable resjobs "Resource job loss effects"
.summ enviro resjobs
```

Variable	Obs	Mean	Std. Dev.	Min	Max
enviro	734	-1.38e-09	.9112633	-1.307359	2.209436
resjobs	734	9.40e-11	.9039373	-1.912148	.9766908

علماء الاجتماع يقومون بإنشاء متغيرات مركبة جديدة والبحث عن دليل عن صدق هذه المتغيرات أو علاقتها مع ما نعتقد أن هذه المتغيرات تقيس، في هذا المثال هناك نوع واحد من الصدق، وهو الصدق الظاهري والذي يُدعم بواسطة نمط قابل لشرح التشبع العاملي. النوع الثاني وهو الصدق المعياري الذي يمكن اكتشافه باختبار ما إذا كانت المتغيرات الجديدة مرتبطة مع متغيرات أخرى التي توقعتها دراسات أو نظريات سابقة، فمثلاً أصدق

نتيجة من بحث استقصائي حول قضايا البيئة في الولايات المتحدة كانت نتيجة انتشار تأثيرات أيديولوجية أو خط حزب معين، الرسم البياني الصندوقي يعرض توزيع القيم على عامل القضايا البيئية (*enviro*) مع الانتماء السياسي للمشاركين في الدراسة *party*، حيث إن الرسم يعرض بوضوح هذا النمط المتوقع، في الشكل (7.11) هناك خط أفقي يمثل الوسيط العام والذي تم الحصول عليه من نتيجة **r(p50)** الخاصة بالأمر **summarize, detail** مع أوزان معاينة الاستقصاء التي تُستخدم كأوزان تحليلية، ضمن المشاركين في الدراسة الذين يعرّفون أنفسهم بأنهم جمهوريون لهم قيم عالية للمتغير *enviro*، وهذه القيم تظهر وكأنها قيم متطرفة.

```
.quietly summ enviro, detail
.graph box enviro [aw = surveywt], over(party)
 yline(`r(p50)')
```



الشكل (9.11)

النتيجة الأكثر شيوعاً – والتي كانت ضمن نتائج البحوث السابقة – تتضمن تأثيرات للعمر والجنس والتعليم. باحثو CERA مهتمون حول ما إذا

كان الإدراك البيئي أو التوعية البيئية للسكان الجدد بالمناطق الريفية يختلف عن إدراك السكان الذين يعيشون منذ فترة طويلة. الانحدار المعروض أدناه وجد تأثيرات ذات معنوية للتعليم، والانتماء السياسي، والقادمين الجديد، المشاركون في الدراسة الذين لهم مستوى تعليم أعلى وهم إما ديمقراطيون أو مستقلون وهم أيضاً عاشوا في المنطقة لأكثر من 5 سنوات في العادة يتفهّمون التأثيرات المحلية الناتجة عن المشاكل البيئية.

```
. svy: regress enviro age-newcomer
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =         1          Number of obs      =        734
Number of PSUs     =       734          Population size    = 725.97798
                                        Design df          =        733
                                        F(   5,     729)   =      15.64
                                        Prob > F           =     0.0000
                                        R-squared          =     0.1275
```

enviro	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	-.0008836	.0027041	-0.33	0.744	-.0061923	.004425
sex	.1037294	.0807127	1.29	0.199	-.0547262	.2621849
educ	.1037535	.0383315	2.71	0.007	.0285008	.1790062
party	-.2949115	.0430073	-6.86	0.000	-.3793436	-.2104794
newcomer	-.2197846	.1089288	-2.02	0.044	-.4336341	-.005935
_cons	.2841083	.2264291	1.25	0.210	-.1604185	.7286351

هناك انحدار مشابه مع العامل المتعلق بمصادر الوظائف *resjobs* والذي وجد بأن هذا المتغير أقل تأثراً بالتعليم أو السياسة، وبدلاً من ذلك فإن متغير العمر والسكان الجدد هما أقوى متغيرات تنبؤية. فصغار السن المشاركون في الدراسة والذين انتقلوا للمنطقة خلال السنوات الخمس الماضية في العادة أقل إدراكاً للتأثيرات الناتجة من خسارة الوظائف في مجال الغابات وصيد الأسماك.

```
. svy: regress resjobs age-newcomer
```

```
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =         1                Number of obs      =        734
Number of PSUs     =       734                Population size    = 725.97798
                                              Design df          =        733
                                              F(   5,     729)   =       6.02
                                              Prob > F           =     0.0000
                                              R-squared          =     0.0792
```

resjobs	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.008514	.0031023	2.74	0.006	.0024235	.0146045
sex	.0826771	.0883132	0.94	0.349	-.0906998	.2560541
educ	.0613762	.0414013	1.48	0.139	-.0199031	.1426555
party	-.0835506	.0479571	-1.74	0.082	-.1777002	.010599
newcomer	-.3624841	.1330725	-2.72	0.007	-.6237327	-.1012354
_cons	-.488574	.2391453	-2.04	0.041	-.9580654	-.0190826

القياس ونماذج المعادلة الهيكلية :

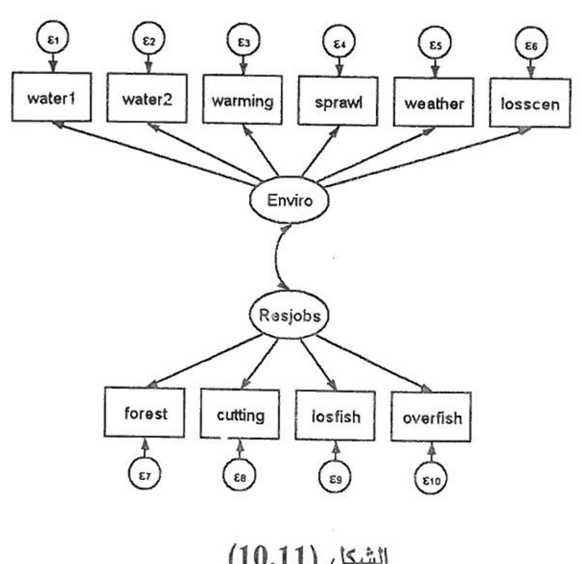
Measurement and Structural Equation Models

الفصل (8)، ألقى نظرة على نماذج المعادلة الهيكلية بداية من مثال الانحدار، الذي يتضمن علاقات بين مجموعة متغيرات (الشكل 14.8)، نماذج المعادلة الهيكلية يمكن أن تشمل نماذج قياس، والتي تشبه التحليل العاملي. نماذج القياس تشير إلى واحد أو أكثر من المتغيرات الكامنة العاملية أو غير المشاهدة والتي تسبب تبايناً في المتغيرات المشاهدة، الشكل (10.11) يوضح ذلك باستخدام بيانات استقصاء منطقة شمال غرب المحيط الهادئ CERA.

```
.use C:\data\PNWsurvey2_11.dta, clear
```

يظهر في الشكل (10.11) متغيران كامنان وغير مشاهدين وهما *Enviro*، *Resjobs*. أسماء هذه المتغيرات تم قصداً وضعهما بحيث تشبه القيم العاملية في الجزء السابق من هذا الفصل، ولكن في هذا التحليل سوف نبدأ من جديد، فلا توجد متغيرات لها نفس هذه الأسماء؛ عند صياغة نموذج

المعادلة الهيكلية يقوم برنامج ستاتا (وهذا هو الوضع الافتراضي للبرنـــامج) باتباع طريقة لتمثيل المتغيرات الكامنة مع أسماء تبدأ بحروف كبيرة، المتغير الكامن *Enviro* يتم رسمه بحيث يوضح أن بعض التباينات في ستة متغيرات غير مشاهدة، ونفس هذه المتغيرات يتم تحميلها بشكل رئيس للعامـــل الأول factor 1 في الجزء السابق، المتغير الكامن *Resjobs* يشرح التباين في أربعـــة متغيرات غير مشاهدة أخرى، والتي يتم تحميلها بـــشكل رئـــيس أو جزئـــي للعامل الثاني factor 2، لاحظ بأن العرض البياني للأشكال الدائرية يوضـــح المتغيرات الكامنة، بينما المستطيلات تعرض المتغيرات المشاهدة، والـــسهم المائل ذو الرأسين يمثل العلاقة غير السببية بين المتغيرين *Enviro*، *Resjobs*.



الشكل (10.11)

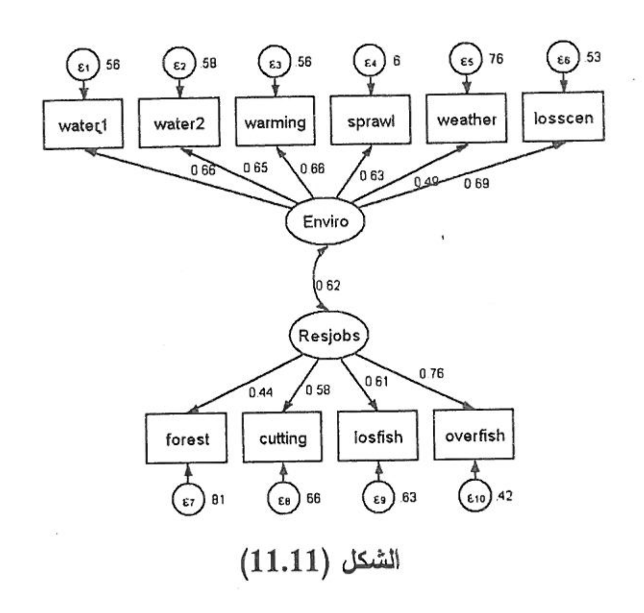
تم إنشاء الشكل أعلاه باستخدام SEM Builder ببرنامج ســـتاتا، ويمكنـــك القيام بذلك عن طريق الخطوات التالية:

Statistics > SEM (structural equation modeling) > Model building and estimation

اختر أداة Add measurement Component (M) من الهامش الأيسر للنافذة، ثم ضع مؤشر الفارة في الموقع الذي تريده للمتغير الكامن بمنطقـــة الرسم، قم بإعطاء المتغير الكامن اسماً يبدأ بحرف كبير مثـــل *Enviro*، ثـــم

اختر متغيرات القياس Measurement variables ليتم شـــرحها بواســطة المتغير الكامن، وذلك باختيار أسمائها من قائمة، بعد ذلك قم باختيار اتجــاه القياس Measurement direction مثل UP واضغط OK، ثم قم بإعادة هذه الخطوات مع المتغير الكامن الثاني في موقع آخر مثل *Resjobs* مـــع وضـــع Measurement direction إلى Down، ثم اضغط OK مرة أخرى، أخيراً قم باستخدام أداة Add covariance (C) من الهامش الأيسر وذلـــك لوضـــع السهم المنحني ذي الرأسين للتباين أو الارتباط بين المتغيرات الكامنة.

الشكل (11.11) يعرض نفس نموذج القياس بعد التقدير، المُعَاملات مـــع مساراتها من المتغيرات الكامنة إلى المتغيرات المـــشاهدة تمثـــل مُعَـــاملات الارتباط المعيارية، وهي تشبه التشبعات العاملية. كل متغير مشاهد له قيمـــة تساوي قيمة تباينه الفريدة والتي تُعطي بواسطة شروط □ (إبـــسلون epsilon) في نموذج المسار هذا، المتغيران *Enviro* و *Resjobs* لهما ارتباط يساوي 0.62.



الشكل (11.11)

عند رسم نموذج مسار مثل الذي يظهر في الشكل (10.11) قم بـــالنقر على Estimate وسوف ترى أن النتائج الإحصائية بدأت بـــالظهور علـــى الرسم البياني. برنامج ستاتا يعرض معلومات كثيرة افتراضياً، ولكن لبعض الأغراض – مثل أغراض النشر – قد نحتاج إلى الاحتفاظ بالرسم بـــشكل

أبسط، الشكل (11.11) يتضمن أوزان الاستقصاء، ومُعامِلات الانحدار المعيارية أو التباينات. ويتم عرض كل ذلك في تنسيق ثابت مع رقمين في يمين الفاصلة العشرية، وتم تبسيط الرسم البياني والتحكم فيه من خلال خيارات القوائم التالية:

Settings > Variables > All ... > Results > Exogenous variables > None > OK

Settings > Variables > All ... > Results > Endogenous variables > None > OK

Settings > Variables > Error ... > Results > Error std. variance > OK

Settings > Connections > Paths > Results > Std. parameter > OK

Settings > Connections > All > Results > Result 1 > Format %3.2f > OK > OK

Estimation > Estimate > Weights > Sampling weights ... (قم باختيار متغيرات مثل surveywt) > OK

نفس النموذج يمكن تقديره مباشرة عن طريق الأمر **sem** بدون استخدام SEM Builder، في الأمر أدناه لاحظ استخدام :svy قبل الأمر والتي تقوم بتطبيق وزن للاستقصاء للأمر **sem** كما قامت بفعله من قبل مع الأمر **regress** والعديد من عمليات ستاتا الأخرى، المُعامِلات المعيارية من هذه المخرجات تتوافق مع مُعَامِلات المسار في الشكل (11.11)، كما أن التباينات الفريدة الخاصة بالمتغيرات المشاهدة والتغاير المعياري (أي الارتباط) بين المتغيرات الكامنة، وبالمثل فإنها تتوافق مع القيم المعروضة في الشكل (11.11).

```
. svy: sem (Enviro ->water1 water2 warming
  sprawl weather losscen) (Resjobs ->forest
  cutting losfish overfish), standard
(running sem on estimation sample)

Survey: Structural equation model

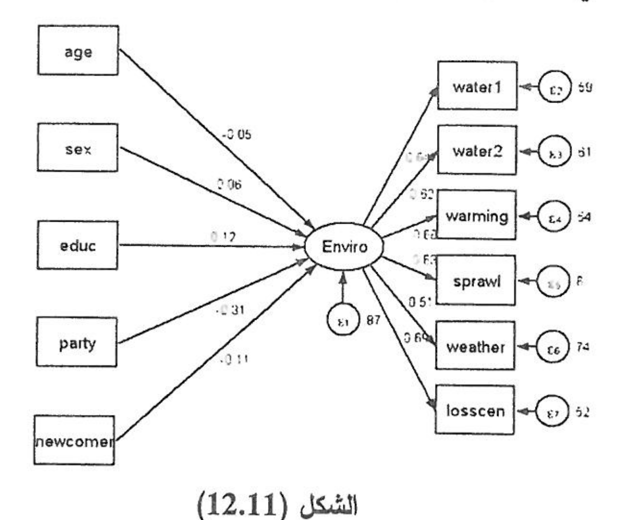
Number of strata   =          1          Number of obs      =       734
Number of PSUs     =        734          Population size    = 725.97798
                                         Design df          =       733

 ( 1)  [water1]Enviro = 1
 ( 2)  [forest]Resjobs = 1
```

Standardized	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf. Interval]	
Measurement						
water1 <-						
Enviro	.6617216	.0305711	21.65	0.000	.6017043	.7217389
_cons	.9865716	.0405518	24.33	0.000	.9069601	1.066183
water2 <-						
Enviro	.6474885	.033683	19.22	0.000	.5813618	.7136151
_cons	.6092422	.0301459	20.21	0.000	.5500596	.6684249
warming <-						
Enviro	.6604354	.0295415	22.36	0.000	.6024393	.7184315
_cons	.9570574	.0398204	24.03	0.000	.8788818	1.035233
sprawl <-						
Enviro	.6331502	.0394133	16.06	0.000	.5557738	.7105265
_cons	.8070991	.0351607	22.95	0.000	.7380715	.8761267
weather <-						
Enviro	.4872326	.0408934	11.91	0.000	.4069505	.5675148
_cons	1.229424	.0493018	24.94	0.000	1.132634	1.326213
losscen <-						
Enviro	.6868967	.0336976	20.38	0.000	.6207413	.7530521
_cons	.8308351	.0357022	23.27	0.000	.7607443	.9009259
forest <-						
Resjobs	.4351381	.0782853	5.56	0.000	.2814481	.5888282
_cons	1.701551	.0851467	19.98	0.000	1.534391	1.868712
cutting <-						
Resjobs	.5800483	.0483413	12.00	0.000	.4851445	.6749521
_cons	1.086461	.0455767	23.84	0.000	.9969847	1.175938
losfish <-						
Resjobs	.611518	.0646856	9.45	0.000	.4845268	.7385092
_cons	1.699434	.0899735	18.89	0.000	1.522798	1.876071
overfish <-						
Resjobs	.7640356	.0396915	19.25	0.000	.686113	.8419582
_cons	1.211127	.0531622	22.78	0.000	1.106759	1.315495

Variance						
e.water1	.5621246	.0404591			.4880516	.6474398
e.water2	.5807587	.0436187			.5011403	.6730265
e.warming	.563825	.0390206			.4921959	.6458784
e.sprawl	.5991209	.049909			.5087318	.7055699
e.weather	.7626044	.0398492			.6882511	.8449901
e.losscen	.5281729	.0462936			.4446786	.6273445
e.forest	.8106548	.0681298			.6873535	.9560745
e.cutting	.663544	.0560805			.5620954	.7833023
e.losfish	.6260458	.0791129			.4884978	.8023236
e.overfish	.4162496	.0606515			.3126948	.5540985
Enviro	1	.			.	.
Resjobs	1	.			.	.
Covariance						
Enviro						
Resjobs	.6194808	.0603082	10.27	0.000	.5010833	.7378782

الشكل (12.11) يعرض مثالاً يتم فيه دمج نموذج قياس (متغير كامن *Enviro* موضحاً التباين في ستة متغيرات مشاهدة، وهي من المتغير *water1* وحتى المتغير *losscen*) مع نموذج انحدار يكون فيه المتغير *Enviro* نفسه يتم تفسيره بواسطة خمسة متغيرات خلفية مشاهدة، وهي من متغير *age* وحتى متغير *newcomer*، وبالتالي فإن الشكل (12.11) يشبه التحليل الذي قمنا به في الجزء السابق آخذاً القيم العاملية للمتغير *enviro* كمتغير تابع له انحدار على المتغيرات من المتغير *age* وحتى المتغير *newcomer*، طريقة المعادلة المركبة تقوم بدمج سمات التحليل العاملي والانحدار في نموذج واحد، والطريقة تقوم بتقدير واختبار مدى واسعاً من الطرق الأخرى لصياغة النماذج مثل ارتباطات الأخطاء، والعلاقات التي تتضمن المتغيرات الكامنة أو المتغيرات المشاهدة الأخرى.



الشكل (12.11)

الأمر التالي يقوم بتقدير نفس النموذج الذي شاهدناه في الشكل (12.11) مع تفاصيل أكثر.

```
.svy: sem (Enviro ->water1 water2 warming
 sprawl weather losscen) (age sex educ party
 newcomer ->Enviro), standard
 (running sem on estimation sample)

Survey: Structural equation model

Number of strata   =      1          Number of obs      =       734
Number of PSUs     =    734          Population size    = 725.97798
                                     Design df          =       733

 ( 1)  [water1]Enviro = 1
```

Standardized	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
Structural						
Enviro <-						
age	-.049815	.0575978	-0.86	0.387	-.1628914	.0632614
sex	.0613274	.051095	1.20	0.230	-.0389826	.1616375
educ	.1203573	.0491607	2.45	0.015	.0238449	.2168698
party	-.3127998	.0500495	-6.25	0.000	-.4110573	-.2145423
newcomer	-.1093366	.0540813	-2.02	0.044	-.2155094	-.0031639
Measurement						
water1 <-						
Enviro	.6428964	.0331036	19.42	0.000	.5779073	.7078856
_cons	1.310619	.1795706	7.30	0.000	.9580847	1.663153
water2 <-						
Enviro	.6245504	.0376486	16.59	0.000	.5506384	.6984623
_cons	.9240407	.1715633	5.39	0.000	.5872266	1.260855
warming <-						
Enviro	.6783106	.0304815	22.25	0.000	.6184691	.7381521
_cons	1.298954	.1864988	6.96	0.000	.9328181	1.665089
sprawl <-						
Enviro	.6309674	.0405903	15.54	0.000	.5512803	.7106546
_cons	1.125132	.1758644	6.40	0.000	.7798741	1.47039
weather <-						
Enviro	.5098352	.0405605	12.57	0.000	.4302065	.5894638
_cons	1.486401	.1467739	10.13	0.000	1.198254	1.774549
losscen <-						
Enviro	.6949899	.034121	20.37	0.000	.6280033	.7619764
_cons	1.181138	.1923889	6.14	0.000	.8034391	1.558837

Variance				
e.water1	.5866842	.0425644	.5087998	.6764906
e.water2	.6099368	.0470269	.5242611	.7096139
e.warming	.5398947	.0413519	.4645213	.6274983
e.sprawl	.6018801	.0512223	.5092718	.7113288
e.weather	.7400681	.0413584	.6631687	.8258846
e.losscen	.5169891	.0474275	.4317822	.6190105
e.Enviro	.8653113	.0330933	.8027214	.9327815

وجدنا بأن المتغيرات *educ, party, newcomer* جميعها لها تـــأثيرات ذات معنوية إحصائية على المتغير الكامن *Enviro*، بينما المتغيرين *age, sex* ليس لهما هذا التأثير. هذه النتائج تتوافق مع النتائج السابقة لانحدار القيم العامليـــة *enviro* على نفس متغيرات التنبؤية الخمسة.

موضوع نماذج المعادلة الهيكلية مع برنامج ستاتا هو موضوع واســـع يستحق كتاباً خاصاً له وحده. الأمثلة في هذا الفصل وتلك التي تم تناولها في الفصل (8) كانت عبارة عن نظرة سريعة. وللحصول علـــى شـــرح كامـــل للأوامر قم بالاطلاع على دليـــل المـــستخدم *Structural Equation Modeling Reference Manual*، كما يحتوي هذا الدليل على نحو مكتبة بهـــا 26 مثـــالاً تسهّل عملية الفهم للقارئ.



الفصل الثاني عشر

تحليل السلاسل الزمنية

Time Series Analysis

قدرات برنامج ستاتا مع السلاسل الزمنية تم تغطيتها في 700 صــفحة بدليل المستخدم *Time-Series Reference Manual*. هذا الفصل يعطـــي مقدمــة مختصرة تبدأ مع أداتين تحليليتين أساسيتين هما: الرسومات البيانية للـــزمن، والتمهيد. ثم ننتقل إلى توضيح استخدام تصوير الارتباط، ونمـــاذج ARIMA وARMAX معاً، مــع الاختبـــارات التشخيـــصية للاســـتقرارية stationarity والضجة البيضاء white noise. هناك تطبيقات أخرى وأهمها شكل الذبذبات، ومجموعة نماذج ARCH المرنة تم تركها للقارئ للبحث عنها واستكشافها.

وبالرغم من أن الحلول التقنية لموضوعات الـــسلاسل الزمنيـــة يمكـــن العثور عليها في دراسة Hamilton (1994)، وهناك مصادر أخـــرى تتـــضمن أيـــضاً دراســـة Box, Jenkins and Reinsel (1994), Chatfield (2004), Diggle (1990), Enders (2004) and Shumway (1988).

قوائم عمليات تحليل السلاسل الزمنية تأتي تحت العناوين التالية:

Statistics > Time series

Statistics > Multivariate time series

Statistics > Cross-sectional time series

Graphics > Time-series graphs

Example Commands : أمثلة عن الأوامر

```
.ac y, lags(8) level(95) generate(newvar)
```

يقوم بإنشاء رسم بياني للارتباطات الذاتية للمتغير *y* مع فترات ثقة 95% (هذا هو الوضع الافتراضي للبرنامج) لفترات تباطؤ من 1 وحتى 8، ويقوم بحفظ الارتباطات الذاتية كأول 8 قيم للمتغير *newvar*.

```
.arch D.y, arch(1/3) ar(1) ma(1)
```

يقوم بتوفيق ARCH (الانحدار الذاتي لاختلاف التباين المشروط autoregressive conditional heteroskedasticity) نموذج مع أول اختلافات للمتغير *y*، وهذا يتضمن شروط ARCH من الدرجة الأولى وحتى الدرجة الثالثة، والدرجة الأولى لأخطاء AR وMA.

```
.arima y, arima(1,1,1)
```

يقوم بتوافق عينة مع نموذج ARIMA(1,1,1) مع اختلاف الدرجة الأولى وشروط اختلاف الدرجة الأولى AR وMA، الخيارات المحتملة الأخرى يمكنها تحديد استراتيجيات تقدير بديلة والقيود الخطية والتقديرات الموثوقة للتباين.

```
.arima y, arima(1,0,2) sarima(1,0,1,12)
```

يقوم بتوافق نموذج $ARIMA(1,0,2)\times(1,0,1)_{12}$ مع شروط AR من الدرجة الأولى، وشروط MA من الدرجتين الأولى والثانية، وكذلك مع المكونات الموسمية المضاعفة مع فترة 12.

```
.arima y x1 x2 x3, arima(2,0,1)
```

يقوم بحساب انحدار ARMAX (المتوسط المتحرك للانحدار الذاتي مع المتغيرات الخارجية) للمتغير *y* على ثلاثة متغيرات تنبؤية. الأخطاء تم صياغتها كمتوسط متحرك من الدرجة الأولى، وانحدار ذاتي من الدرجة الثانية.

```
.arima D.y x1 L1.x1 x2, ar(1) ma(1 12)
```

يقوم بتوافق نموذج ARMAX والذي فيه اختلافات أولية للمتغير *y* تم حساب انحدارها على المتغير *x1* مع قيم لها فترات تباطؤ 1- للمتغير *x1* والمتغير *x2* وهذا يتضمن أخطاء AR(1), MA(1), MA(12).

```
.corrgram y, lags(8)
```

يقوم بحساب الارتباطات الذاتية، والارتباطات الذاتية الجزئية، واختبارات Q لفترات تباطؤ من 1 وحتى 8 للمتغير y.

```
.dfuller y
```

يقوم بحساب اختبار جذر وحدة ديكي فولر Dickey-Fuller للاستقرارية stationarity.

```
.estat dwatson
```

بعد الأمر **regress** مع بيانات سلاسل زمنية يقوم الأمر أعلاه بحساب إحصائية دوربن واتسون Durbin-Watson لاختبار الارتباط الذاتي من الدرجة الأولى.

```
.egen newvar = ma(y), nomiss t(7)
```

يقوم بإنشاء متغير جديد *newvar* يساوي فترة 7- ناقلاً متوسط المتغير y ومستبدلاً قيم البداية والنهاية بمتوسطات أقصر وغير مركزية.

```
.generate date = mdy(month,day,year)
```

يقوم بإنشاء متغير *date* يساوي عدد الأيام منذ 1 يناير 1960 من ثلاثة متغيرات هي *month, day, year*.

```
.generate date = date(str_date, "mdy")
```

يقوم بإنشاء متغير *date* من متغير نصي *str_date*، حيث إن *str_date* يحتوي على تواريخ يكون ترتيبها شهر ويوم وسنة ويكون تنسيقها مثل "1948, June, 12" أو "98/4/18" أو "2001/11/19"، ولمعرفة دوال التاريخ، والخيارات الأخرى، قم بطباعة الأمر **help dates**.

```
.generate newvar = L3.y
```

يقوم بإنشاء متغير جديد *newvar* يساوي قيم المتغير y مع فترة تباطؤ 3.

```
.pac y, lags(8) yline(0) ciopts (bstyle(outline))
```

يقوم بإنشاء رسم بياني للارتباطات الذاتية الجزئية مع فترات ثقة وتباين متبقي من فترة تباطؤ 1 وحتى فترة تباطؤ 8، ويقوم بإنشاء خط أفقي عند نقطة 0، ويقوم الأمر بعرض فترة الثقة كخط عريض بدلاً من منطقة مضللة (والأخير هو الوضع الافتراضي).

```
.pergram y, generate(newvar)
```

يقوم برسم عينة ذبذبات (دالة كثافة الطيف) للمتغير *y* ، وإنشاء متغير جديد باسم *newvar* يساوي قيم الذبذبات الخام.

```
.smooth 73 y, generate(newvar)
```

يقوم بإنشاء متغير جديد *newvar* يساوي الفترة الزمنية 7- مستخدماً قيم الوسيط للمتغير *y*، ثم يقوم بإعادة التمهيد باستخدام الفترة الزمنية 3- مستخدماً قيم الوسيط، التمهيد المركب مثل "3RSSH" أو "4253h,twice" تكون محتملة، وللحصول على معلومات أكثر عن المُرشّحات والتمهيدات قم بطباعة الأمر **help smooth** أو **help tssmooth**.

```
.tsset date, format(%td)
```

يقوم باعتبار البيانات وكأنها بيانات سلاسل زمنية، حيث تم الإشارة إلى الوقت باستخدام المتغير *date* والذي يكون تنسيقه على أساس يومي، أما بالنسبة للبيانات الطويلة panel data مع السلاسل الزمنية المتوازية لعدد من الوحدات المختلفة مثل المدن، فإن الأمر **tsset *city year*** يقوم بتحديد متغيرات الزمن والطول للبيانات، أغلب الأوامر في هذا الفصل تتطلب استخدام الأمر **tsset** مع البيانات.

```
.tssmooth ma newvar = y, window(2 1 2)
```

يقوم بتطبيق المُرشّح المتوسط المتحرك للمتغير *y* مؤدياً إلى إنشاء متغير جديد باسم *newvar*، الخيار **window(2 1 2)** وجد بأن المسافة الزمنية 5- للمتوسط المتحرك وذلك من خلال إدراج فترات تباطؤ تساوي 2 وإدراج المشاهدة الحالية و2 من القيم الأساسية في حساب كل نقطة تمهيدية. وللحصول على قائمة المُرشّحات الأخرى المتوافرة والمتوسطات المتحركة الموزونة والأسيّة والأسيّة المزدوجة وهولت – وينترس Holt-Winters وغير الخطيّة، قم بطباعة الأمر **help tssmooth**.

```
.tssmooth nl newvar = y, smoother(4253h,twice)
```

يقوم بتطبيق مرشح التمهيد غير الخطي على المتغير *y* مؤدياً إلى إنشاء متغير جديد *newvar*، الخيار **smoother(4253h, twice)** يقوم بشكل تكراري بإيجاد قيم الوسيط لمسافات زمنية هي 4، 2، 5، 3 ثم تطبيق دالة هانج

Hanning وبعد ذلك يقوم بتكرار العملية مع البواقي. الأمر **tssmooth nl** يختلف عن باقي أوامر **tssmooth**، حيث لا يمكنه العمل مع القيم المفقودة.

```
.wntestq y, lags(15)
```

يقوم باحتساب اختبار Liung-Box portmanteau Q للضجة البيضاء white noise (كما يمكن حساب هذا الاختبار باستخدام الأمر **corrgram**).

```
.xcorr x y, lags(8) xline(0)
```

يقوم بإنشاء رسم بياني للارتباط المتقاطع بين مدخلات (x)، ومخرجات (y) لفترات تباطؤ من 1–8. الخيار **xcorr *x y*, table** يعطي نسخة نصية تتضمن الارتباطات الفعلية. وإذا قمنا بإضافة الخيار **generate(*newvar*)** إلى الأمر **xcorr** فسوف يتم حفظ الارتباطات لفترات التباطؤ من 8– وحتى 8+ كمتغير في أول 17 صفاً من البيانات.

التمهيد : Smoothing

العديد من السلاسل الزمنية يكون لها تباينات عالية التردد تجعل من الصعب توضيح الأنماط الكامنة، تمهيد smoothing مثل هذه السلاسل تقوم بتقسيم البيانات إلى جزءين، الأول يتباين بشكل تدريجي، والثاني "تقريباً" يحتوي على التغيرات الكبيرة المتبقية:

البيانات = التمهيد + التقريب

ولتوضيح طرق التمهيد سوف نقوم باختبار بيانات عن الاستهلاك اليومي للمياه لقرية ميلفورد Milford بولاية هامبشير Hampshire بالولايات المتحدة خلال فترة سبعة أشهر من يناير وحتى يوليو 1983 (البيانات بالملف *MILwater.dta*، المصدر: دراسة Hamilton 1985)، الأنماط المعتادة لاستخدام المياه في ميلفورد كانت قد تغيرت بواسطة الأخبار المقلقة التي حدثت في منتصف فترة الدراسة.

```
.describe
```

```
Contains data from C:\data\MILwater.dta
  obs:           212                          Milford daily water use, 1/1/83 - 7/31/83
 vars:             4                          2 Jul 2012 06:11
 size:         1,272

              storage  display     value
variable name   type   format      label      variable label

month           byte   %9.0g                  Month
day             byte   %9.0g                  Date
year            int    %9.0g                  Year
water           int    %9.0g                  Water use in 1000 gallons

Sorted by:
```

قبل التعمّق في التحليل، سوف نحتاج إلى تحويل معلومات الشهر واليوم والسنة إلى مؤشر رقمي واحد للزمن. دالة برنامج ستاتا ( )mdy تقوم بهذا التحويل، حيث تقوم بإنشاء متغير الوقت الماضي (والذي سوف نطلق عليه هنا *date*) مشيراً إلى عدد الأيام منذ 1 يناير 1960.

```
.generate date = mdy(month,day,year)
.list in 1/5
```

	month	day	year	water	date
1.	1	1	1983	520	8401
2.	1	2	1983	600	8402
3.	1	3	1983	610	8403
4.	1	4	1983	590	8404
5.	1	5	1983	620	8405

التاريخ المرجعي وهو 1 يناير 1960 هو تاريخ عشوائي ولكن ثابتاً، يمكننا أن نجعل تنسيق التاريخ أكثر فهماً للمتغير *date*، كما يمكننا تجهيز البيانات للتحليل لاحقاً، وذلك باستخدام الأمر tsset (وهو اختصار للحروف الأولى من كلمة بيانات سلاسل زمنية time series set) لتحديد التاريخ *date* كمتغير لمؤشر الزمن، وتحديد الخيار td(daily)% لعرض هذا المتغير.

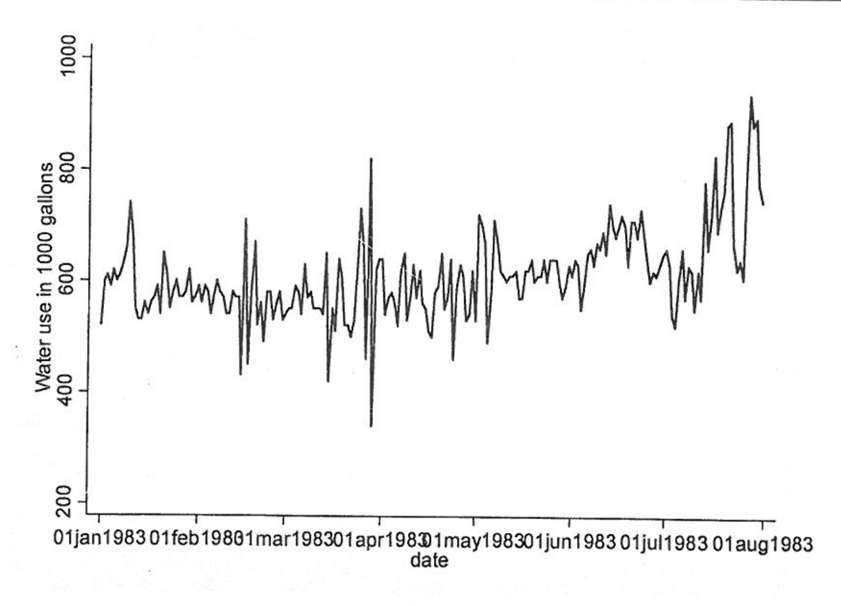
```
.tsset date, format(%td)
 time variable:  date, 01jan1983 to 31jul1983
 delta:  1 day
.list in 1/5
```

	month	day	year	water	date
1.	1	1	1983	520	01jan1983
2.	1	2	1983	600	02jan1983
3.	1	3	1983	610	03jan1983
4.	1	4	1983	590	04jan1983
5.	1	5	1983	620	05jan1983

التواريخ في المتغير الجديد *date* لها تنسيق مثل "05 jan 1983"، وهو أكثر فهماً وقراءة عن القيم الرقمية المجردة مثل "8405" (وهو عبارة عن عدد الأيام منذ 1 يناير 1960)، كما يمكننا استخدام التنسيق **%td** لإنتاج تنسيقات أخرى مثل "05 Jan 1983" أو "83/05/01". برنامج ستاتا يوفر عددًا من المتغيرات التعريفية، وعددًا من تنسيقات العرض وتنسيقات البيانات التي لها أهميتها مع السلاسل الزمنية. العديد من هذه التنسيقات يتضمن طرقاً للمدخلات والتحويلات وعرض التواريخ. ويمكنك الحصول على شرح مفصل عن دوال التواريخ في دليل المستخدم *Data Management Reference Manual* ودليل المستخدم *User's Guide* أو يمكنك الاطلاع عليها بطباعة الأمر **help dates**.

الشكل (1.12) يستخدم شكلاً بيانياً من نوع **twoway line** لرسم شكل بياني بسيط لمتغير استخدام المياه *water* مع الزمن *date*. الشكل يعرض نمط تباين يوم بيوم، كما يوضح أن هناك ارتفاعاً في استخدام المياه في الصيف، قيم متغير التاريخ تم وصفها تلقائياً (01 jan 1983 وهكذا) على المحور الأفقي *x* ، ولكن خيارات ستاتا الافتراضية هنا تقود إلى نتائج غير مرضية، وتوصيفات متزاحمة في الرسم.

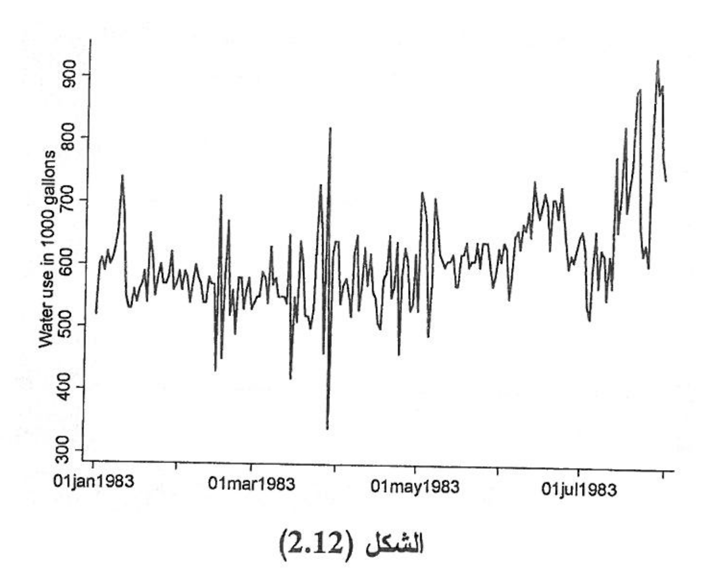
```
.graph twoway line waterdate
```



الشكل (1.12)

هناك طريقة أفضل لرسم الزمن، واستخدام متغير التاريخ على المحور الأفقي *x*، ويتم ذلك باستخدام أمر خاص بالسلاسل الزمنية وهو **tsline**، هـذا الأمر يسمح لنا بشرح قيم المحور *x* في ضوء التاريخ بدون الحاجـة إلـى التواريخ في أسفل المحور. فمثلاً يمكننا رسم شكل يوضح الوقت يشبه الشكل (1.12) ولكن توصيفات المحور الأفقي لن تكون متزاحمة كما فـي الـشكل (1.12). لاحظ أن الأمر **tsline** لا يقبل أي متغير من متغيرات *x*، فهو يقبل فقط واحد أو أكثر من متغيرات *y*، مع بيانات **tsset** أن سمة الوقت تم تعريفها مسبقاً، والخيارات **tlabel( )** و**ttick( )** تعمل في الرسم البيـاني **tsline**، كمـا تعمل الخيارات **xlabel( )** و**xtick( )** مع أي شكل بياني مـن النـوع **twoway** باستثناء أنها تفهم تنسيقاً معيناً للوقت مثل 01 jan 1983، في الـشكل (2.12) قمنا باختصار عناوين المحور الأفقي *x* (الوقت أو محور *t*) باستخدام الخيار **ttitle(" ")** لأن كلمة تاريخ "Date" تبدو غير ضرورية أسفل المحور الأفقي الذي تم توصيفه 01 jan 1983، 01 mar 1983 وهكذا.

```
.tsline water, ylabel(300(100)900) ttitle("")
   tlabel(01jan1983 01mar1983 01may1983
   01jul1983, grid)
   ttick(01feb1983 01apr1983 01jun1983
   01aug1983)
```



الشكل (2.12)

الفحص بالعين المجردة يلعب دورًا رئيساً في السلاسل الزمنية. فالتمهيد يساعدنا في أن نرى الأنماط الكامنة تحت السلاسل المتذبذبة. طريقة التمهيد الأبسط يتم استخدامها لحساب المتوسط المتحرك عند كل نقطة زمنية بناءً على القيم الحالية والسابقة واللاحقة للمتغير y. فمثلاً "المتوسط المتحرك للفترة الزمنية 3" يُشير إلى المتوسط y_{t-1} ، y_t ، y_{t+1}، ويمكننا استخدام مميزات ستاتا لإنشاء **generate** مثل هذا المتغير:

```
.generate water3a = (water[_n-1] + water[_n] +
 water[_n+1])/3
```

الطريقة الأفضل تتضمن ma (المتوسط المتحرك) كدالة للأمر egen:

```
.egen water3b = ma(water), nomiss t(3)
```

الخيار **nomiss** في الأمر **egen** يجعل المتوسطات المتحركة أقصر وغير مركزة في أطراف المنحنى، إذا لم نقم باستخدام هذا الخيار، فإن أول وآخر قيمة للمتغير *water3* سوف تكون قيماً مفقودة. الخيار **t(3)** يجعل المتوسط

المتحرك يكون للفترة الزمنية 3، وأي عدد إضافي للفترة الزمنية 3 أو أكثر يمكن استخدامه.

هناك أدوات تمهيد قوية متوافرة لبيانات السلاسل الزمنية (tsset) ويتم استخدامها من خلال الأوامر **tssmooth**، كلها يمكنها التعامل مع البيانات المفقودة باستثناء **tssmooth nl**

tssmooth ma	مُرشّحات المتوسط المتحرك أو الموزونة أو غير الموزونة.
tssmooth exponential	المُرشّحات الأسيّة الفردية.
tssmooth dexponential	المُرشّحات الأسيّة المزدوجة.
tssmooth hwinters	تمهيد هلت – وينترز Holt-Winters غير الموسمية.
tssmooth shwinters	تمهيد هلت – وينترز Holt-Winters الموسمية.
tssmooth nl	المُرشّحات غير الخطية.

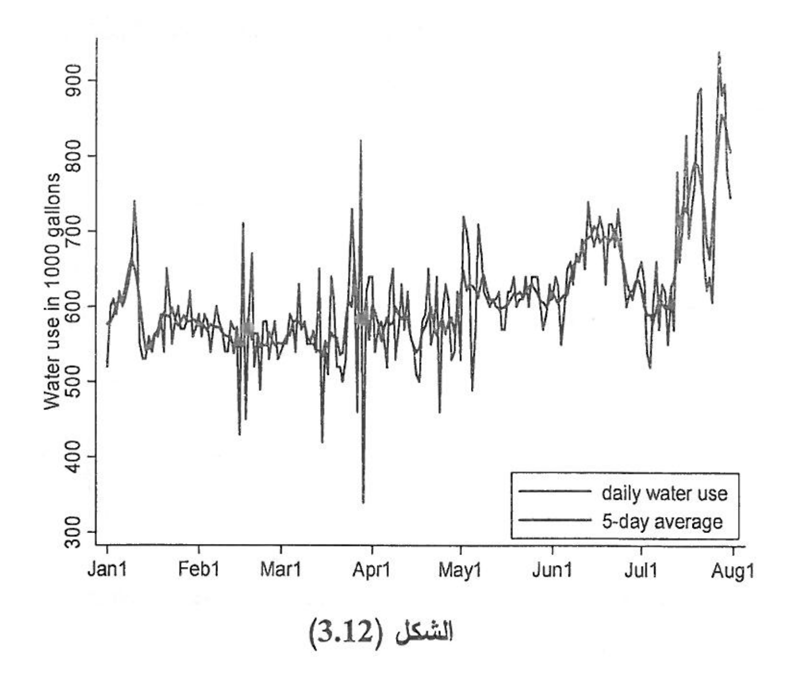
فعلى سبيل المثال، فإن الأمر **tssmooth ma** يمكنه حساب المتوسطات المتحركة للفترة الزمنية 3، وهي مطابقة لتلك الناتجة من أمر **egen** السابق:

```
.tssmooth ma water3c = water, window(1 1 1)
The smoother applied was
(1/3)*[x(t-1) + 1*x(t) + x(t+1)]; x(t)= water
```

لمعرفة تركيبة كل أمر، قم بطباعة الأمر **help tssmooth ma** أو الأمر **help tssmooth exponential**.

الشكل (3.12) يرسم المتوسط المتحرك لخمسة أيام لاستخدام المياه في قرية ميلفورد (*water5*) معاً مع البيانات الخام (*water*). الأمر **tsline** يقوم بتركيب رسم الفترة الزمنية لقيم المتغير *water5* الممهّد على الرسم البياني للقيم الخام للمتغير *water* (خط أقل سُمكاً). توصيفات المحور الأفقي تم إنشاؤها باستخدام التواريخ كما فعلنا سابقاً مع الشكل (2.12). في الشكل (3.12) قمنا بتحديد تنسيق عرض مبسط للتاريخ معطياً فقط الشهر واليوم **format(%tdmd)**، هذا التنسيق المبسط للتاريخ يترك مسافة لتوصيف بداية كل شهر في هذا الرسم خلافاً لكل شهر آخر، كما تم سابقاً في الشكل (2.12).

```
.tssmooth ma water5 = water, window(2 1 2)
The smoother applied was
(1/5)*[x(t-2)+ x(t-1)+ 1*x(t)+x(t+1)+x(t+2)]; x(t)=water
.tsline water, clwidth(medium)
    || tsline water5, clwidth(medthick)
    || , ylabel(300(100)900) ytitle("Water use
    in 1000 gallons")
    ttitle("") tlabel(01jan1983 01feb1983
    01mar1983 01apr1983
    01may1983 01jun1983 01jul1983 01aug1983,
    grid format(%tdmd))
    legend(position(4) ring(0) rows(2)
    label(1 "daily water use") label(2 "5-day
    average"))
```



الشكل (3.12)

المتوسطات المتحركة تشترك مع الإحصائيات الأخرى التي تعتمد على المتوسط في عيب واحد، وهو أنها أقل مقاومة للقيم المتطرفة. وحيث إن القيم المتطرفة واضحة من الحواف المدببة البارزة في السلاسل الزمنية للمياه بالشكل أعلاه، لذلك فإننا قد نقوم بمحاولة استخدام طريقة تمهيد أكثر مقاومة للقيم المتطرفة. الأمر **tssmooth nl** يقوم بتمهيد غير خطي مقاوم للقيم

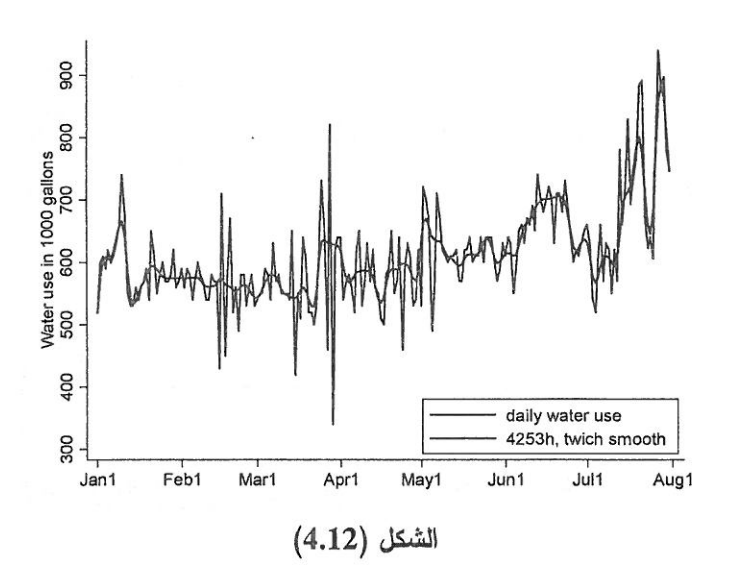
المتطرفة مستخدماً طرقاً ومصطلحات تم شرحها في دراسة Velleman and Hoaglin (1981) ودراسة Velleman (1982)، فعلى سبيل المثال:

```
.tssmooth nl water5r = water, smoother(5)
```

يتم إنشاء متغير جديد باسم *water5r* يحتوي على قيم المتغير *water* بعد تمهيدها بواسطة استخدام قيم الوسيط لفترة زمنية 5، المُمهّدات المركبة باستخدام قيم الوسيط المتعلقة بالفترات الزمنية مع مجموعة دالة هانج (¼، ½، ¼–) للمتوسط المتحرك الموزون لفترة زمنية تساوي 3) وتقنيات أخرى يمكنها تحديد الرمز الأصلي لفالمان Velleman، الممهّد المركب الواحد الذي يبدو مفيدًا مع البيانات التي تتغير بسرعة يُطلق عليه "4253h, twice" وبتطبيق هذا المُمهّد على متغير *water* يمكننا حساب المتغير الممهد *water4r*.

```
.tssmooth nl water4r = water, smoother
 (4253h,twice)
```

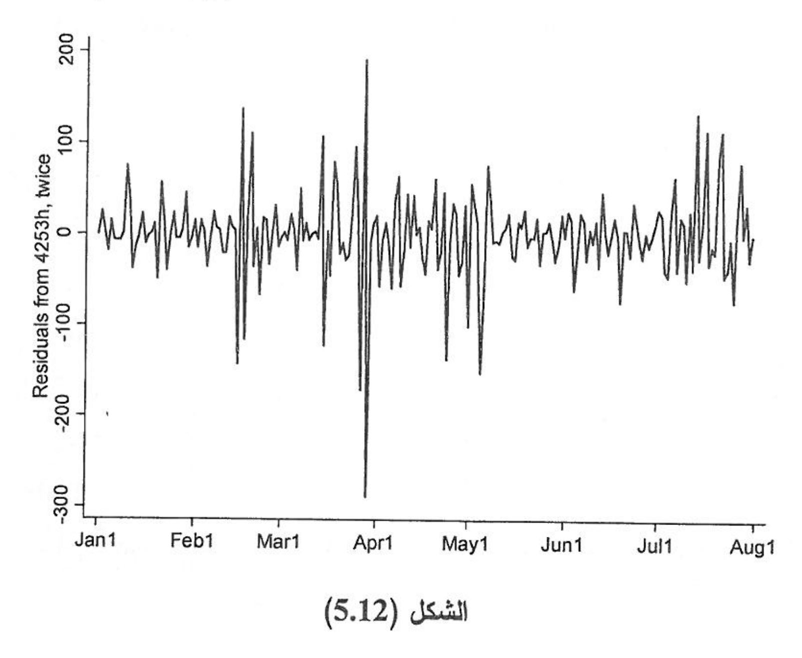
الشكل (4.12) يعرض المتغيرات الممهدة الجديدة *water4r* ، وبمقارنة الشكل (4.12) مع الشكل (3.12) نرى كيف أن الممهد 4253h, twice كان أداؤه مقارباً للمتوسط المتحرك لفترة زمنية تساوي 5–، وبالرغم من أن كلا التمهيدين لهما فترات زمنية متشابهة، فإن الخيار **4253h, twice** قام بتخفيض أكثر للتباينات المدببة بالرسم.



الشكل (4.12)

في بعض الأحيان، هدفنا الرئيسي من التمهيد، هو البحث عن نمط في الرسم البياني للتمهيد. ومع هذه البيانات خصوصاً فإن البواقي أو قاع السلسلة بعد التمهيد يكون أكثر إثارة. ويمكننا أن نحسب القاع كفرق بين البيانات والتمهيد، ثم نقوم بتمثيل البواقي بيانياً مع فتراتها الزمنية، كما فعلنا سابقاً في الشكل (5.12).

```
.generate rough = water - water4r
.label variable rough "Residuals from 4253h,
 twice"
.tsline rough, ttitle("")
 tlabel(01jan1983 01feb1983 01mar1983 01apr1983
 01may1983 01jun1983 01jul1983 01aug1983, grid
 format(%tdmd))
```



الشكل (5.12)

أقوى التقلبات في الشكل (5.12) حدثت في الفترة ما بين 27-29 مارس. استخدام المياه انخفض فجأة ثم ازداد مرة أخرى، وبعد ذلك عاد لينخفض بشكل أقل، ثم ارتفع بشكل أكبر قبل الاستقرار في مستوياته المعتادة. في هذه الفترة الصحف المحلية كانت تتحدث عن النفايات الكيميائية

الخطيرة التي تم اكتشافها في واحد من الآبار التي تقوم بتزويد القرية بالماء. التقارير الأولية حذّرت الناس، وانخفض استهلاك المياه بشكل ملحوظ. خلال الأيام التالية استخدام المياه تراوح ما بين أعلى وأقل قمة في المنحنى، وذلك كرد فعل للتطورات الجديدة في الاستهلاك للاستخدام في الفتـرة الأخيـرة، وعادت الأشياء للاستقرار من جديد بعد أن تم استبعاد البئر موضع التساؤل من الخدمة.

أمثلة أكثر عن الرسومات البيانية للزمن :

Further Time Plot Examples

البيانات الموجودة بالملف *Greenland_temperature.dta* تتضمن سلـسلة زمنية معروفة لتقديرات درجة الحرارة والتي تم إعادة إنشائها مـن GISP2 جليد جرينلاند، حيث تغطي البيانات فترة زمنية تصل إلـى 50,000 سـنة ماضية حتى سنة 1855 (دراسة Alley 2004). في المنشورات العلمية عـن هذا النوع من البيانات، يتم تمثيل الزمن باستخدام متغير *age* الذي يـستخدم وحدات تمثل آلاف السنوات الماضية. "الوقت الحالي" تم تعريفـه بواسـطة الباحثين في مجال الجليد بأنه يعني سنة 1950، وأحدث البيانات الموجـودة بمتغير *age* هي 0.095 أو 95 سنة ماضية، أو بعبارة أخـرى هـي سـنة 1855. الجليد والثلوج للسنوات الأخيرة لم تكن متراكمة بما فيه الكفاية حتى نطبّق عليها طريقة إعادة تركيب درجة الحرارة. ولجعل بيانات مثالنا أكثـر وضوحاً لغير المتخصصين في علم المناخ، فسوف يتم إنشاء متغيـر جديـد للوقت باسم *year*، حيث يمكن قراءة هذا المتغيـر "كتقـويم سـنوي" مـن 48,000- إلى 1999، المتغير *gisptemp* يتضمن درجات الحرارة التي تـم إنشاؤها من بيانات الجليد GISP2. المتغير *shutemp* يحتوي علـى متوسـط درجات الحرارة السنوية في نفس الموقع بجزيرة جرينلاند، والتي قام العلماء بقياسها خلال الفترة من 1987-1999 (دراسة Shuman et al. 2001).

```
.use C:\data\Greenland_temperature.dta, clear
.describe
```

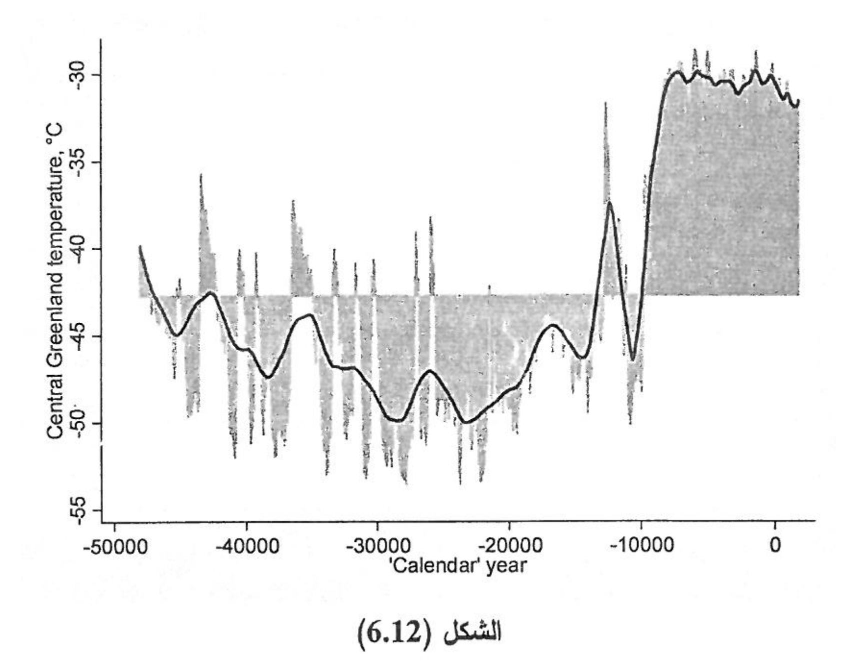
```
Contains data from C:\data\Greenland_temperature.dta
  obs:         1,646                          Greenland ice core temp 48,000 years ago to
                                                1855 (Alley 2004)
 vars:             4                          2 Jul 2012 06:11
 size:        26,336
--------------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
--------------------------------------------------------------------------------------
year            float  %5.0f                  'Calendar' year
age             float  %8.0g                  Age, 1,000s of years before 1950
gisptemp        float  %8.0g                  GISP2 ice core Central Greenland temp -48,000 to
                                                1855, C
shutemp         float  %9.0g                  Shuman (2001) Summit temp 1987 to 1999, C
--------------------------------------------------------------------------------------
Sorted by:  year
```

الشكل (6.12) يعرض الرؤوس المدببة لبيانات درجات الحرارة GISP2. هذه الرؤوس المدببة تم إنشاؤها من مستوى أساسي يساوي المتوسط للأمد الطويل مع درجات الحرارة التي تكون أعلى من المتوسط والتي تم الإشارة إليها في الرسم باستخدام رؤوس مدببة حمراء، ودرجات الحرارة التي تكون أقل من المتوسط والتي تم الإشارة إليها في الرسم برؤوس مدببة زرقاء. وحيث إن البيانات أكثر توافراً للسنوات الأخيرة، فإن المستوى الأساسي المناسب لكامل السلسلة تم حسابه من متوسط فترات آلاف السنين بدلاً من متوسط القياسات الفردية.

```
.gen millennium = int(year/1000)
.collapse (mean) gisptemp, by(millennium)
.summ gisptemp
```

Variable	Obs	Mean	Std. Dev.	Min	Max
gisptemp	50	-42.75702	6.798864	-51.48554	-30.02251

```
.use C:\data\Greenland_temperature.dta, clear
.graph twoway spike gisptemp year if gisptemp>
 -42.75702,base(-42.75702) lcolor(erose)
    || spike gisptemp year if gisptemp<= -
       42.75702,
    base(-42.75702) lcolor(eltblue)
    || lowess gisptemp year, bwidth(.05)
       lwidth(medthick)lcolor(black)
    || , ytitle("Central Greenland temperature,
       `=char(176)'C")legend(off)
```



الشكل (6.12)

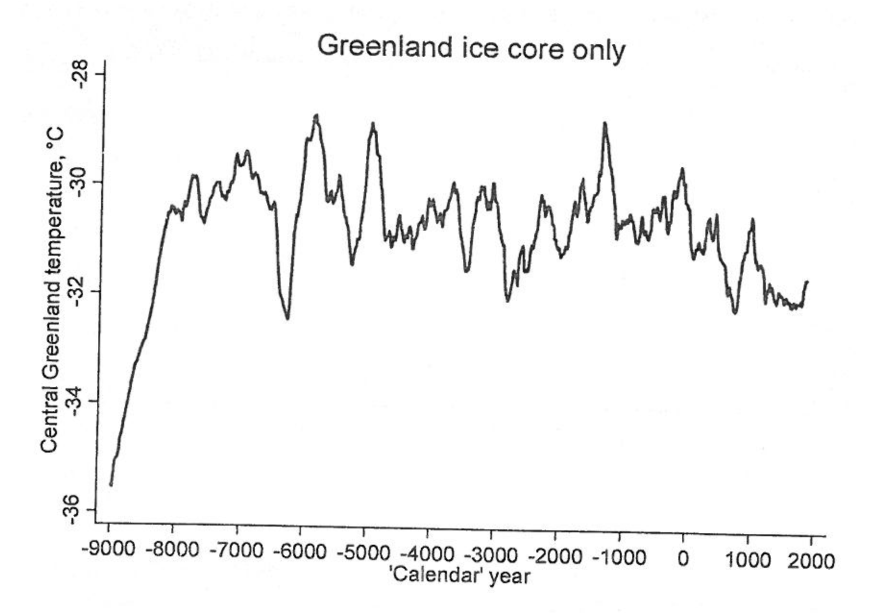
هذا الشكل البياني يعرض بوضوح عملية التحول من العصور الجليدية إلى الظروف الأكثر حرارة. درجات الحرارة في جرينلاند زادت بحوالي 20 درجة مئوية، من أقل من 50- درجة مئوية إلى حوالي 30- درجة مئويـــة، هذا التحول توقف مؤقتاً حيث تمت العودة من جديد للعصور الجليديـــة فـــي الفترة التي تسمى Younger Dryas والتي حدثت في الفترة ما بـــين 12,900 سنة تقريباً و11,500 سنة ماضية (التقويم السنوي 10,900- إلى 9,500- في الرسم البياني، الرأس المدبب المنخفض الأخير)، بداية ونهاية Younger Dryas كل منها حدثت خلال عقود قليلة أو أقل من ذلك، دافعاً العديـــد مـــن الدراسات لاختبار السبب المحتمل والإمكانية المـــستقبلية للتغيـــر المنـــاخي المفاجئ (انظر على سبيل المثال، دراسة White et al. 2010).

المتوسط المتحرك أو تقنيات التمهيد غير الخطية التي تم توضيحها في الجزء السابق، تعمل بشكل جديد مع المشاهدات التي توجد بينها فترات زمنية متساوية، وعندما يحدث تباين بين هذه المشاهدات في فترات زمنية قصيرة، فإن قياسات الجليد في GISP2 بها مسافات زمنية متفاوتة بين مشاهداتها، مما

يجعل الفترة الزمنية بين مشاهداتها متفاوتة أكثر كلما تعمقنا أكثر (أقدم) جاعلاً طبقات الجليد أكثر كثافة. بالنسبة للسلاسل الزمنية مع فترات زمنية متفاوتة أو التي تم إجراء تمهيد لها خلال فترات زمنية طويلة أو التي لها انحدار لدالة تمهيد مثل المنحنى الذي يظهر في الشكل (6.12) فإنها تعتبر بديلاً عملياً.

الشكل (7.12) يعرض أحدث جزء من البيانات التي تظهر في الشكل (6.12) واصفاً درجات الحرارة GISP2 لفترة 11,000 سنة الماضية فقط، الأشكال البيانية المشابهة لهذا الشكل البياني الذي يعرض الزمن بيانياً كان مصدراً أساسياً لعدم الوضوح.

```
.graph twoway line gisptemp year if year> -
 9000,lwidth(medthick)
 xlabel(-9000(1000)2000, grid gmax gmin)
 title("Greenland ice core only")
 ytitle("Central Greenland temperature,
 `=char(176)'C")
```



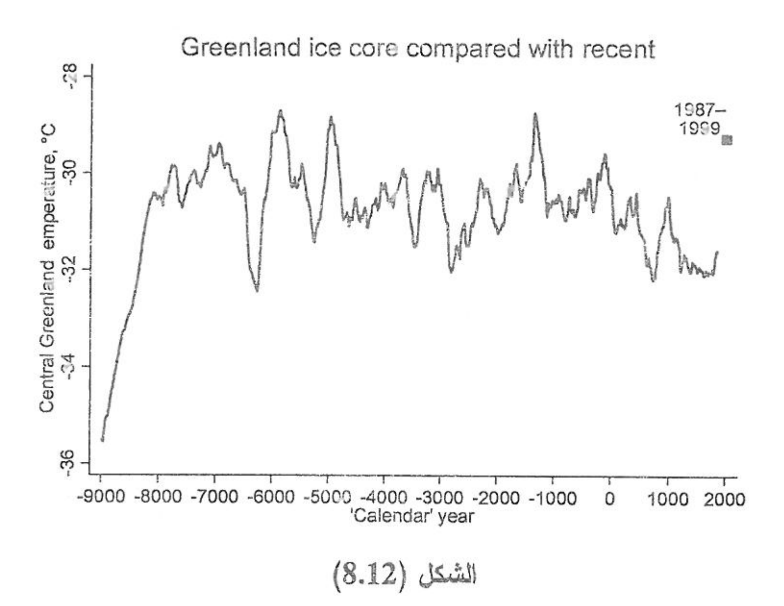
الشكل (7.12)

يمكننا أن نرى أن العديد من قيم التباين الخادعة في الشكل (7.12) حيث إن نقطة البيانات النهائية (الواقعية 1855) تم توصيفها بأنها "الحالية"، وإذا

كانت هذه النقطة تمثل درجة الحرارة الواقعية، فإننا قد نعتقد بأن جرينلانــد مازالت أكثر برودةً من المتوسط مقارنة بـــ 10,000 ســـنة الماضـــية، وأن الارتفاع الذي حدث أخيراً في درجات الحرارة كان غير ملحوظ، وهذه هي النقطة التي حاول هذا الشكل إظهارها. البعض حاول استخدام طريقة الرسم البياني باستخدام خط أساسي مشابه لما قمنا به في الشكل (6.12) ولكن تــم تحديد 1855 على أنها نقطة النهاية للخط الأساسي، وذلك لجعل"البـــرد الآن" رسالة أكثر وضوحاً. المتغيرات الأخرى تُخفي حقيقة أن درجات الحرارة في جرينلاند ليست درجات الحرارة العالمية.

في 1855 عموماً، فإن درجات الحرارة في جرينلاند كانت تزداد بدرجة بسيطة جداً من فترة برودة يُطلق عليها العصر الجليدي البـــسيط. درجـــات الحرارة الحديثة أكثر دفئاً، وهذا أدى إلى انخفاض في كتلة الغطاء الجليدي، كما أدى إلى انخفاض في جليد البحر حول الساحل. حتى في فترة التسعينيات القياسات المباشرة من قمة الغطاء الجليدي توضح أن متوسط درجة الحرارة السنوية هو 29.26− درجة مئوية (دراســـة Shuman et al. 2001)، الـــشكل (8.12) يُعيد رسم الشكل (7.12) ولكن مع نقطة بيانـــات نهائيـــة تعـــرض درجات الحرارة للفترة ما بين 1987−1999 يتم تركيبها كنقطة في الرســـم البياني. هناك خطان من النصوص تم وضعهما في المركز عند محور y عند درجة الحرارة 28.80−، وعند التوصيف 1500 بالمحور الأفقي x وبـــنفس لون العلامة نفسها.

```
.graph twoway line gisptemp year,
 lwidth(medthick)
    || scatter shutemp year, msymbol(S)
       mcolor(red)
    || if year >- 9000,xlabel(-9000(1000)2000,
       grid gmax gmin)
   title("Greenland ice core compared with
   recent") legend(off)
   ytitle("Central Greenland temperature,
   `=char(176)'C")
   text(-28.80 1500 "1987`=char(150)'" "1999",
   color(red))
```



الشكل (8.12)

الشكل (8.12) يترك انطباعاً مختلفاً عن الانطباع الذي يتركه الشكل (7.12)، بالرغم من أن بيانات الجليد المستخدمة في كلا الشكلين هي نفسها.

التغيرات الأخيرة في المناخ : Recent Climate Change

سوف ننقل تركيزنا على وحدات القياس من آلاف السنوات إلى ثلاثين سنة ماضية فقط. بقية هذا الفصل، سوف تركز على التغيرات التي حدثت في المناخ خلال السنوات الأخيرة. ملف البيانات *Climate.dta* يحتوي على ثلاث سلاسل زمنية تقوم بتقدير درجات الحرارة العالمية الشهرية في الفترة من 1980 وحتى 2010، مع أربعة محركات أو أسباب محتملة لدرجات الحرارة، اثنان من مؤشرات درجات الحرارة يتم استخراجهما من قياسات درجة حرارة السطح (NASA وNCDC) بينما المؤشر الثالث يقوم بتقدير درجات الحرارة من طبقة التربوسفير السفلية والتي تم الحصول عليها من بيانات الأقمار الصناعية (UAH). انظر إلى ملحق مصادر البيانات لمزيد من المعلومات عن كل هذه المتغيرات والتي تم جمعها معاً من مصادر مختلفة.

```
.use C:\data\Climate.dta, clear
.describe
```

```
Contains data from C:\data\Climate.dta
  obs:          372                          Global temperature & drivers 1980-2010
 vars:           11                          2 Jul 2012 06:11
 size:       13,764

              storage  display     value
variable name   type   format      label      variable label

year            int    %9.0g                  Year
month           byte   %9.0g                  Month
myear           int    %tmmCY                 Month, year
ncdctemp        float  %9.0g                  NCDC global temp anomaly v.1901-2000, C
nasatemp        float  %8.0g                  NASA global temp anomaly v.1951-1980, C
uahtemp         float  %9.0g                  UAH global temp anomaly v.1981-2010, C
aod             float  %8.0g                  Aerosol Optical Depth at 550nm
tsil            float  %8.0g                  Total Solar Irradiance, W/m2
mei             float  %9.0g                  Multivariate ENSO Index
co2globe        float  %8.0g                  Global average marine surface CO2, ppm
co2anom         float  %9.0g                  Global CO2 anomaly, ppm

Sorted by:  myear
```

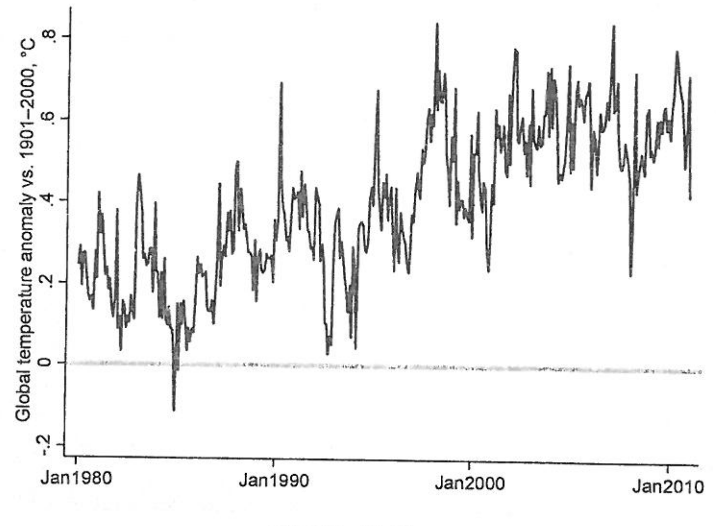
الفترة الزمنية 1980-2010 تم تغطيها بالبيانات الموجودة بالملف *Climate.dta* وهي محددة بالبيانات التي تم الحصول عليها للدراسة وتتضمن: سلسلة زمنية لمتوسط مستويات ثاني أكسيد الكربون تم استخدامها هنا من بداية سنة 1980، وبيانات العمق البصري للهباء الجوي Aerosol Optical Depth والتي تم تحديثها حتى سنة 2010. بيانات هذا الملف تم تحديدها كبيانات سلاسل زمنية باستخدام الأمر **tsset** مستخدماً المتغير *myear* كمتغير يمثل الزمن، وهو متغير يحتوي على بيانات شهرية تم تعريفها من خلال فصل المتغير *month* والمتغير *year* والتي كانت موجودة في البيانات الأصلية. وداخلياً يقوم برنامج ستاتا بتعريف المتغيرات الشهرية كرقم لعدد الأشهر منذ يناير 1960 بنفس الطريقة التي تم بها تعريف متغيرات التاريخ كرقم لعدد الأيام منذ 1 يناير 1960، هنا المتغير *myear* له تنسيق **%tmmCY** حتى يظهر لنا 1980 Jan، 1980 Feb وهكذا. إذا لم يتم القيام بإجراء هذا التنسيق مسبقاً فإن إنشاء هذا المتغير الشهري، وتحديد البيانات كبيانات سلاسل زمنية باستخدام الأمر **tsset**، هي خطوات ضرورية للقيام بأي تحليل لاحقاً.

```
.gen myear = ym(year, month)
.format myear %tmmCY
.label variable myear "Month, year"
.tsset myear
```

في ورقة بحثية تم استخدامها كمرجع من قبل كثير مــن الدراســات، قام كل من Foster وRahmstorf 2011 (بنــاءً على عمــل ســابق من قبــل 2008 Lean and Rind,) بتحليل درجات حرارة مشابهة، ومتغيــرات قياديــة للكشف عن "الإشارات الحقيقية لظاهرة الاحتباس الحراري العالمي"، التحليل الذي قاموا به أكثر دقة من الأمثلة البسيطة الموجودة في هذا الفــصل مــع وصولنا لنتائج متشابهة في العموم.

الشكل (9.12) يعرض درجات الحرارة العالمية الشاذة حسب مؤشر مركــز بيانات المناخ الوطني (NCDC) والتي تظهر فــي المتغيــر *ncdctemp*، درجــات الحرارة الشاذة من محطة أرصاد جوية تمثل انحرافات عــن درجــة الحــرارة المشاهدة من متوسط درجات الحرارة في الأمد الطويل في محطة الأرصاد تلك، درجات الحرارة الشاذة العالمية بالتالي تم حسابها من متوسط موزون للعديد مــن درجات الحرارة الشاذة في العديد من محطات الأرصــاد حــول العــالم، الخــط الرمادي عند درجة الحرارة صفر يمثل متوسط القرن العشرين.

```
.tsline ncdctemp, lw(medthick)
 ytitle("Global temperature anomaly vs.
 1901`=char(150)'2000,
 `=char(176)'C", size(medsmall))
 yline(0, lcolor(gs12) lwidth(thick))
 xtitle("")
 xlabel(, grid gmin gmax)
```



الشكل (9.12)

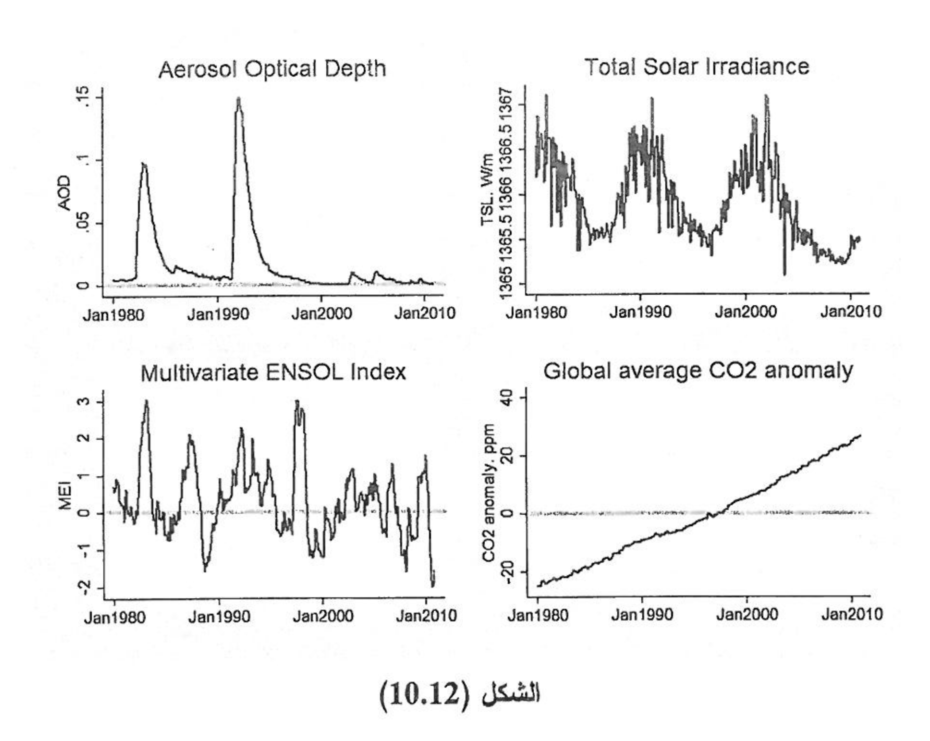
كل الأشهر ماعدا شهرين في الشكل (9.12) لها درجات حرارة أعلى من متوسط درجات الحرارة في القرن العشرين. ظاهرة الاحتباس الحراري شهدت ارتفاعاً كبيراً منذ منتصف السبعينيات، وهي الفترة التي تم تمثيلها بيانياً في هذا الشكل، والنقطة المركزية العليا في هذا الشكل وهو الشهر الذي شهد أعلى درجات حرارة عالمياً في فبراير 1998 لم يتم تجاوزها (ثم تم الوصول إليها تقريباً) حتى يناير 2007. كلتا النقطتين أكثرمن الانحرافات المعيارية للاثنين، وأعلى من أي شهر قبل سنة 1980 في بيانات NCDC والتي تعود إلى سنة 1880 (انظر الشكل 1.2 في الفصل الثاني)، وعموماً فإن درجات الحرارة المرتفعة خلال فترة زمنية كبيرة من سنة 1998 مع درجات الحرارة المنخفضة خلال العقد الماضي تُشكل انطباعاً واضحاً بأن الاحتباس الحراري قد يكون قد توقف مؤقتاً خلال العقد الماضي. الجزء الأخير من هذا الفصل، يقوم بتطبيق نماذج السلاسل الزمنية لاختبار التفسيرات المحتملة.

إذا كانت هناك أسباب تنظيمية، فإنها قد توجد بين المتغيرات الأخرى في بيانات *Climate.dta*. هذه تتضمن أربعة عوامل تم تحديدها بواسطة علماء الطبيعة كعوامل مهمة تسبب أو تؤثر على تباين درجة حرارة الهواء. وحيث إن عدم صفاء الجو والتي تُقاس هنا بواسطة العمق البصري للهباء الجوي (AOD) عند طول موجي 550nm وهذا يعكس تأثير ثورات البراكين والتي تجعل درجة حرارة السطح أكثر برودة، وذلك من خلال إضافة أشعة ضوء الشمس التي تحجزه البراكين المرتفعة في الجو (دراسة Sato et al. 1993). الخفوت الشمسي الكلي Total Solar Irradiance (TSI) يتضمن قياسات تعتمد على الأقمار الصناعية للكمية المتذبذبة من أشعة الشمس (تُقاس بالوات لكل متر مربع) الذي يسطع عند الطبقات العلوية للغلاف الجوي للأرض (دراسة Frohlich 2006)، وحدوث ظاهرة إلنينو/ التذبذب الجنوبي (ENSO) يمكن أن تكون لها تأثيرات جوهرية على درجات حرارة الهواء العالمية، كما أن الهواء الساخن الذي يحدث أثناء تغيرات الجو ويساهم في تسخين سطح الماء في مركز وشرق المنطقة الاستوائية بالمحيط الهادئ (إلنينو El Nino) أو

درجات الهواء البارد، والذي يقوم بدفع مياه المحيط العميقة إلى السطح (لا نينا La Nina)، بعض مؤشرات ENSO تُعرف من خلال درجات حرارة سطح البحر، وتؤدي إلى إنشاء تيارات مدارية، وتستخدم هذه لاحقاً لتفسير درجة الحرارة. ومن ناحية أخرى، فإن مؤشر ENSO المتعدد (MEI) يتم تعريفه من خلال المكوّن الرئيس الأول للمتغيرات 6 التي تم مشاهدتها في المنطقة الاستوائية بالمحيط الهادئ وهي: ضغط مستوى سطح البحر، المكونات الطولية والعرضية للرياح السطحية، درجة حرارة سطح البحر، درجة حرارة هواء السطح، والجزء المُغيّم الكلي من السماء؛ قيم MEI التي تكون أعلى من الصفر بدرجة جوهرية تُشير إلى حالة إلنينو، بينما القيم التي تكون أقل من الصفر تشير إلى لانينا (دراسة Wolter and Timlin 1998)؛ المحرك الرابع في هذه البيانات هو المتوسط العالمي لتركيز ثاني أكسيد الكربون لسطح البحر، ويُقاس بأجزاء من المليون بناءً على قياسات من شبكة المواقع الجغرافية المختلفة (دراسة Masarie and Tans 1995)، المتغير *co2globe* يعطي التركيز الفعلي لثاني أكسيد الكربون، والمتغير *co2anom* يقوم بإزالة التباينات الموسمية، وذلك من خلال طرحها من كل قيمة للمتوسط العالمي لذلك الشهر في الفترة 1980–2010.

الشكل (10.12) يقوم بدمج الرسومات البيانية للزمن لأربع سلاسل رئيسة خلال نفس السنوات التي تم تمثيلها في الشكل (9.12)، الرسم البياني لـ AOD في أعلى اليسار له نمط مثير للانتباه، وتم التأثير عليه بواسطة اثنتين من الثورات البركانية الكبيرة: الأولى الشيكنو El Chichón في المكسيك في نهاية مارس 1982، والثانية بجبل بيناتوبو Mount Pinatubo في الفلبين في يونيو 1991. كما أن TSI يتحرك بشكل متسلسل، ولكن هذا التسلسل تمت مقاطعته فجأة بواسطة رؤوس مدببة مرتفعة ومنخفضة. ثم كانت الفترة الأخيرة مستقرة إلى حد كبير. بالنسبة لـ MEI فلم يكن متسلسلاً بالرغم من مظهره المتذبذب مع فترات غير منتظمة بين الأوضاع الإيجابية والسلبية، عدم الانتظام واحتمال حصول تغيرات سريعة يجعل إلينيو ولانينا أحداثاً صعبة التوقع، بالرغم من تأثيراتها الكبيرة على الناس، مما يجعل التنبؤ بها

هدف مهم جداً. كما أن تركيز ثاني أكسيد الكربون كان سهل التوقع، لأنه في ارتفاع مستمر بزيادة قدرها 35 بليون طن. حيث إنه يزداد في الهواء الجوي كل سنة نتيجة أنشطة بشرية.



الشكل (10.12)

آخر جزء في هذا الفصل، قام باختبار كيف أن هذه المتغيرات الأربعة معاً تشرح التغيرات الأخيرة في درجات الحرارة، ونتابع هذا التحليل في الجزء التالي الذي سوف يتطرق إلى المفاهيم الأساسية والأدوات.

فترات التباطؤ والسوابق والفروقات :

Lags, Leads and Differences

تحليل السلاسل الزمنية في العادة، يتضمن متغيرات لها فترات تباطؤ أو قيم من فترات ماضية. فترات التباطؤ يمكن تحديدها بواسطة الأقواس. فمثلاً الأمر أدناه يقوم بإنشاء متغير جديد باسم *mei_1* وهو يساوي القيم السابقة لمؤشر ENSO التراكمي:

```
.generate mei_1 = mei[_n-1]
```

أو هناك طريقة أخرى يمكننا استخدامها للقيام بنفس العملية، وذلك من خلال الأمر tsset مع المحدد .L (فترة التباطؤ). فاستخدام محددات فترات التباطؤ أفضل من استخدام طريقة الأقواس. والشيء المهم جداً أن استخدام محدد فترة التباطؤ يأخذ في الاعتبار البيانات الطولية Panel Data، الأمر أدناه يقوم بإنشاء قيم تباطؤ لمدة شهر 1- و2- للمتغير *mei*.

```
.generate mei_1 = L1.mei
.label variable mei_1 "MEI 1-month lag"
.generate mei_2 = L2.mei
.label variable mei_2 "MEI 2-month lag"
.list year month mei mei_1 mei_2 in -5/1
```

	year	month	mei	mei_1	mei_2
368.	2010	8	-1.849	-1.217	-.466
369.	2010	9	-2.037	-1.849	-1.217
370.	2010	10	-1.948	-2.037	-1.849
371.	2010	11	-1.606	-1.948	-2.037
372.	2010	12	-1.58	-1.606	-1.948

كما يمكننا الحصول على نفس القائمة، وذلك من خلال إنشاء متغيرات جديدة عن طريق طباعة الأمر:

```
.list year month mei L1.mei L2.mei in -5/1
```

المحدد .L هو أحد الأدوات التي تُسهّل العمل مع بيانات السلاسل الزمنية، حيث هناك محددات أخرى هي .F (لاحقة)، .D (الفرق)، .S (الفرق الموسمي)، هذه المحددات يمكن طباعتها كحرف صغير أو كبير، فمثلاً F2.*mei* هو نفسه f2.*mei*.

محددات السلاسل الزمنية : Time Series Operators

L.	فترة تباطؤ y_{t-1} (.L1 يعني نفس الشيء).
L2.	فترتا تباطؤ y_{t-2} (وبالمثل .L3 وهكذا، .L(1/4) تعني فترات تباطؤ من .L1 وحتى .L4).
F.	فترة لاحقة y_{y+1} (.F1 تعني نفسي الشيء).
F2.	فترتان لاحقتانy_{y+2} (وبالمثل .F3 وهكذا).

D. الفرق بين الفترة الحالية والفترة السابقة $y_t - y_{t-1}$ (**D1.** تعني نفس الشيء).

D2. فرق الدرجة الثانية $(y_t - y_{t-1}) - (y_{t-1} - y_{t-2})$ (وبالمثل **D3.** وهكذا).

S. الفرق الموسمي $y_t - y_{t-1}$ (وهو نفسه الفرق الناتج من **D.**)

S2. الفرق الموسمي الثاني $(y_t - y_{t-2})$ (وبالمثل **S3.** وهكذا).

في حالة الفروقات الموسمية، فإن المحدد **S12** لا يعني الاختلاف الثاني عشر، ولكنه الاختلاف الأول عند فترة تباطؤ 12. فعلى سبيل المثال، إذا كانت لدينا قيم فعلية لثاني أكسيد الكربون CO_2 العالمية فقط بدلاً من القيم الشاذة، فإنه بالإمكان أن نرى نمطاً موسمياً واضحاً منخفضاً في أغسطس وسبتمبر، ولبعض الأسباب قد نحتاج إلى حساب *S12.co2* والذي قد يكون هو الاختلافات بين يناير 1980 *co2* ويناير 1981 *co2*، وكذلك بين فبراير 1980 *co2* وفبراير 1981 *co2* وهكذا.

محددات فترات التباطؤ يمكن أن تظهر مباشرة في أغلب أوامر التحليل التي تتضمن بيانات **tsset**. المثال أدناه يقوم بحساب انحدار درجات الحرارة العالمية (*ncdctemp*) على مؤشر العمق البصري للشهر الماضي (*aod*) والذي يُفترض نظرياً أن يزيد تأثير البرودة. الانحدار تم حسابه بدون الحاجة إلى إنشاء أي متغيرات تباطؤ جديدة.

```
.regress ncdctemp L1.aod
```

Source	SS	df	MS
Model	1.7221338	1	1.7221338
Residual	11.5699674	369	.031354925
Total	13.2921012	370	.035924598

Number of obs =	371
F(1, 369) =	54.92
Prob > F =	0.0000
R-squared =	0.1296
Adj R-squared =	0.1272
Root MSE =	.17707

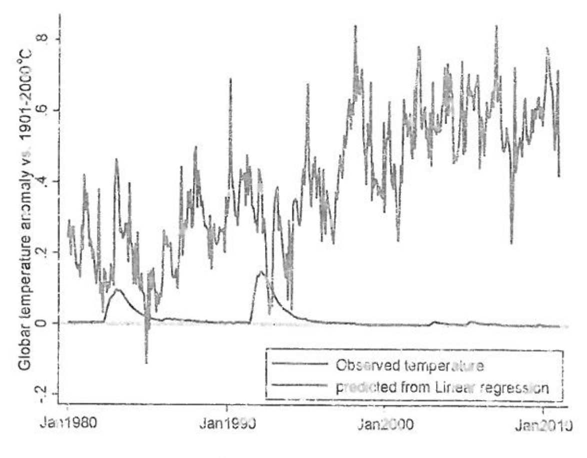
ncdctemp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
aod L1.	-2.313292	.3121404	-7.41	0.000	-2.92709	-1.699495
_cons	.4361341	.0105842	41.21	0.000	.4153213	.456947

وكما هو متوقع، فإن متغير *aod* له تأثير سالب على درجات الحـــرارة العالمية، والنموذج الذي تم تقديره يتضمن درجات حرارة شهرية كدالة لقـــيم *aod* للشهر الماضي:

$$ncdctemp_t = 0.436 - 2.313aod_{t-1}$$ المتوقعة

معامل المتغير المتباطئ *aod*(-2.313) يظهر ذو معنوية إحصائية ($p \approx 0.000$) ولكن الأخطاء المعيارية، واختبارات الانحدار قد لا تكون صـــالحة، وكمـــا هـــو معتاد مع أي نموذج OLS فإنه يعتمد على فرضية أن أخطاء المشاهدات المتعاقبة هي أخطاء مستقلة أو غير مترابطة مع بعضها، الأخطاء المترابطة تحـــدث فـــي العادة في تحليل السلاسل الزمنية. ولذلك فإننا نقوم بشكل روتيني باختبار وجودها كما سوف يتم فعله في الجزء التالي. الأخطاء المعيارية وفترات الثقة واختبارات الانحدار OLS التي تتضمن سلاسل زمنية يجب النظر إليها بنوع من الشك مـــالم تُظهِر الاختبارات بأنه ليس هناك دليل على وجود الأخطاء المترابطة.

بالرغم من احتمالية أن الأخطاء المترابطة تجعل اختبارات *F* و *t* من هذا الانحدار غير موثوقة، فإن معادلة الانحدار نفسها يمكن أن تعطينـــا وصـــفاً صحيحاً للمربعات الصغرى للبيانات، الشكل (11.12) يعرض القيم المتوقعة مع درجات الحرارة المشاهدة، ثورة بركان Pinatubo الـــضخمة فـــي سـنة 1991 توقعت انخفاضاً كبيرًا في درجات الحرارة، والتي تظهر في بيانـــات درجات الحرارة. ومن الواضح أن هناك الكثير مع ثورة بركانين كان لهـــا تأثيرها على المناخ العالمي.



الشكل (11.12)

سوف نستمر في التحليل الوصفي، حيث يمكننا اكتشاف ما إذا كانت محركات درجات الحرارة المقترحة في هذه البيانات تتطور حتى تتناسب مع درجات الحرارة المشاهدة. وبتضمين أشعة الشمس مع فترة تباطؤ كمتغير تنبؤي ثان، فإن ذلك يزيد من قيمة R^2_a زيادة بسيطة من 0.127 إلى 0.144، ويظهر معامل متغير أشعة الشمس التباطؤية سالباً، وهذا ليس له أي معنى منطقي.

```
.regress ncdctemp L1.aod L1.tsi1
```

```
      Source |       SS           df       MS      Number of obs   =       371
-------------+----------------------------------   F(2, 368)       =     32.22
       Model |  1.98067563         2  .990337816   Prob > F        =    0.0000
    Residual |  11.3114256       368   .03073757   R-squared       =    0.1490
-------------+----------------------------------   Adj R-squared   =    0.1444
       Total |  13.2921012       370  .035924598   Root MSE        =    .17532

------------------------------------------------------------------------------
    ncdctemp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         aod |
         L1. |  -2.172641   .3128342    -6.95   0.000    -2.787808   -1.557474
             |
        tsi1 |
         L1. |  -.0598584   .0206393    -2.90   0.004    -.1004441   -.0192727
             |
       _cons |   82.19402   28.19026     2.92   0.004     26.75982    137.6282
------------------------------------------------------------------------------
```

إضافة متغير تنبؤي تباطؤي ثالث وهو مؤشر ENSO المتعدد يزيد من قدرة النموذج على شرح التباين $R^2_a = 0.201$، معامل المتغير L1.*mei* موجب ويُفترض أن يُعطي تأثيرًا حرارياً معروفاً لإلينوي. سوف نرى أيضاً معامل سالب غير معقول للمتغير L1.*tsi1*.

```
.regress ncdctemp L1.aod L1.tsi1 L1.mei
```

```
      Source |       SS       df       MS              Number of obs =     371
-------------+------------------------------           F(  3,   367) =   32.00
       Model |  2.75629902     3  .918766339           Prob > F      =  0.0000
    Residual |  10.5358022   367  .028707908           R-squared     =  0.2074
-------------+------------------------------           Adj R-squared =  0.2009
       Total |  13.2921012   370  .035924598           Root MSE      =  .16943

------------------------------------------------------------------------------
    ncdctemp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         aod |
         L1. |  -2.949131   .3372229    -8.75   0.000    -3.612262   -2.285999
             |
        tsi1 |
         L1. |    -.05509   .0199673    -2.76   0.006    -.0943546   -.0158253
             |
         mei |
         L1. |   .0524371   .0100882     5.20   0.000     .0325991     .072275
             |
       _cons |   75.67739   27.27247     2.77   0.006     22.04748    129.3073
------------------------------------------------------------------------------
```

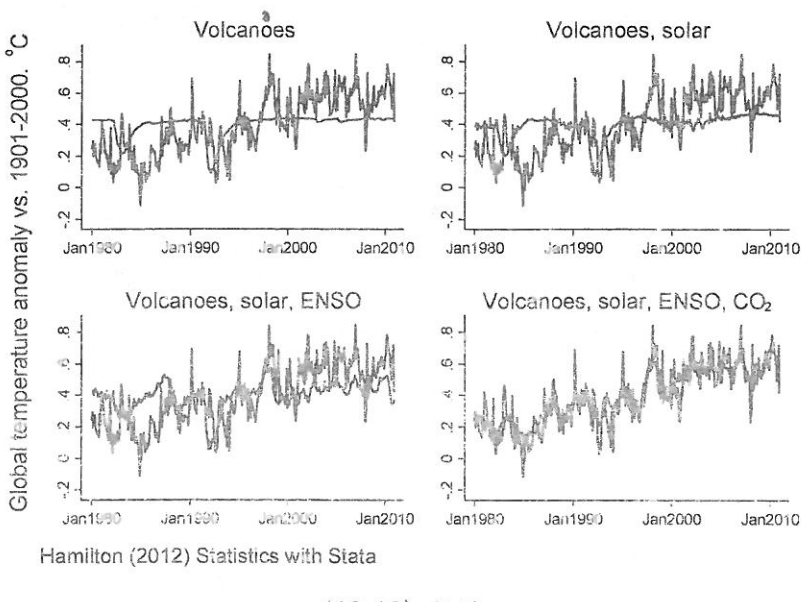
التطور الملحوظ في R^2_a حدث عندما قمنا بإضافة القيم الشاذة لـــ CO_2 إلى المتغيرات التنبؤية. هذه المتغيرات الأربعة معاً تشرح حـــوالي 72.7% من التباين في درجات الحرارة الشهرية. القيم الشاذة لـــ CO_2 لها حتى الآن أقوى تأثير في الاتجاه الموجب كما هو متوقع من طبيعة الغازات المـــسببة للاحتباس الحراري. فعندما نتحكم في CO_2، فإن معامل أشعة ضوء الـــشمس تصبح موجبة أيضاً.

```
.regress ncdctemp L1.aod L1.tsi1 L1.mei
 L1.co2anom
```

```
      Source |       SS       df       MS              Number of obs =     371
-------------+------------------------------           F(  4,   366) =  247.57
       Model |  9.70518563     4  2.42629641           Prob > F      =  0.0000
    Residual |  3.58691559   366  .009800316           R-squared     =  0.7301
-------------+------------------------------           Adj R-squared =  0.7272
       Total |  13.2921012   370  .035924598           Root MSE      =    .099

------------------------------------------------------------------------------
    ncdctemp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         aod |
         L1. |  -1.535808   .2040555    -7.53   0.000    -1.937077    -1.13454
             |
        tsi1 |
         L1. |   .0882862    .012849     6.87   0.000     .0630189    .1135534
             |
         mei |
         L1. |   .0689124   .0059267    11.63   0.000     .0572578    .0805671
             |
     co2anom |
         L1. |   .0109831   .0004125    26.63   0.000      .010172    .0117942
             |
       _cons |  -120.1742   17.55028    -6.85   0.000    -154.6862   -85.66217
------------------------------------------------------------------------------
```

الشكل (12.12) يعرض تطوراً كبيرًا للنماذج الأربعة. القيم المتوقعة في في الجانب الأسفل الأيمن وهي التي ظهرت بناءً على انحدار كل المتغيرات الأربعة الرئيسة يعرض درجات الحرارة المشاهدة بطريقة دقيقة، حيث يتضمن تفاصيل مثل البرودة بعد ثورة بركان Pinatubo، وارتفاع درجة الحرارة خلال فترة "إلنينو الفائق" لسنة 1998 والمسار المتحرك للعقد الماضي، والذي انخفضت درجة حرارته بسبب انخفاض أشعة الشمس والانخفاض الكبير إلى سالب ENSO، التناسب بين درجات الحرارة المتوقعة والمشاهدة واضح جداً، لأن التوقعات جاءت من نموذج بسيط جداً جميع تأثيراته كانت خطية وكانت متغيراته مع فترة تباطؤ لشهر واحد فقط.



الشكل (12.12)

نماذج المناخ العالمي، والتي تعتمد على الطبيعة بدلاً من الإحصاء، تتضمن العديد من المتغيرات، والعديد من التعقيدات الجغرافية المتنوعة والتي قد تحتاج إلى أسابيع لمعالجتها بأجهزة كمبيوتر ذات إمكانات ضخمة جداً. النتائج المعروضة في الشكل (12.12) توضح بأن مثل هذا النموذج البسيط يعمل بقدر إمكاناته وفي ظل محدداته. أحد هذه المحددات هو محدد

إحصائي: حيث إن الأخطاء المعيارية لـ OLS متحيّزة، واختبارات F وt تكون غير صالحة في حالة حدوث ترابط بين هذه الأخطاء. أحد الفحوصات البسيطة للأخطاء المترابطة يُسمى اختبار دربن واتسون Durbin-Watson والذي يمكن استخدامه بعد أي انحدار.

```
.estat dwatson
Durbin-Watson d-statistic(5, 371)= 1.131453
```

كُتب الإحصاء تحتوي على جداول للقيم الحرجة لاختبار دربن واتسون، وإذا أخذنا 5 معلمات مقدّرة (4 متغيرات تنبؤية وتقاطع y) و371 مشاهدة، فإن القيم الحرجة تقريباً هي $d_L = 1.59$ و $d_U = 1.76$، إحصائية الاختبار تحت $d_L = 1.59$ تقود إلى رفض فرضية العدم، بأن هناك ارتباطاً ذاتياً موجباً (فترة تباطؤ 1) من الدرجة الأولى. ولأن الإحصائية المحسوبة 1.131 أقل من $d_L = 1.59$، فإننا يجب أن نرفض فرضية العدم ونستنتج بدلاً من ذلك أن هناك ارتباطاً ذاتياً موجباً من الدرجة الأولى. هذه النتيجة تؤكد الشكوك الأولية حول مدى صلاحية اختبارات في نماذج OLS لدرجات الحرارة.

لو كانت الإحصائية المحسوبة أكبر من $d_U = 1.76$ لفشلنا في رفض فرضية العدم، ولم يكن لدينا أي دليل على أن الارتباط الذاتي ذو معنوية. إحصائيات دربن واتسون المحسوبة بين d_L و d_U فهذا يعني أننا لانستطيع رفض أو قبول فرضية العدم H_0.

إحصائية دربن واتسون تختبر الارتباط الذاتي من الدرجة الأولى، وتأخذ في الاعتبار الفرضيات البديلة الإيجابية فقط. وفي الواقع العملي، فإن الارتباط الذاتي يمكن أن يكون سالباً أو موجباً، ويمكن أن يحدث عند فترات تباطؤ أخرى غير 1. الجزء التالي يعرض بشكل أكثر تفصيلاً أدوات التشخيص العامة.

التمثيل البياني للارتباط : Correlograms

مُعاملات الارتباط الذاتي تُقدّر الارتباط بين متغير ما ونفسه عند فترات تباطؤ معينة، فمثلاً الارتباط الذاتي من الدرجة الأولى هو الارتباط بين y_t

و y_{t-1}، أما الدرجة الثانية فهي تُشير إلى الارتباط بين $[y_t , y_{t-2}]$ وهكذا. التمثيل البياني للارتباط يعرض تمثيلاً بيانياً للارتباط مقابل فترة التباطؤ.

الأمر **corrgram** يعطي تمثيلاً بيانياً مبسطاً للارتباط مع المعلومات ذات العلاقة، أقصى عدد لفترات التباطؤ التي يتم عرضها يمكن أن يكون محدوداً نتيجة البيانات، وذلك باستخدام الخيار **matsize**، أو بعض الأرقام المنخفضة العشوائية، والتي يتم تحديدها باستخدام الخيار **lags( )**:

```
.corrgram mei, lags(13)
```

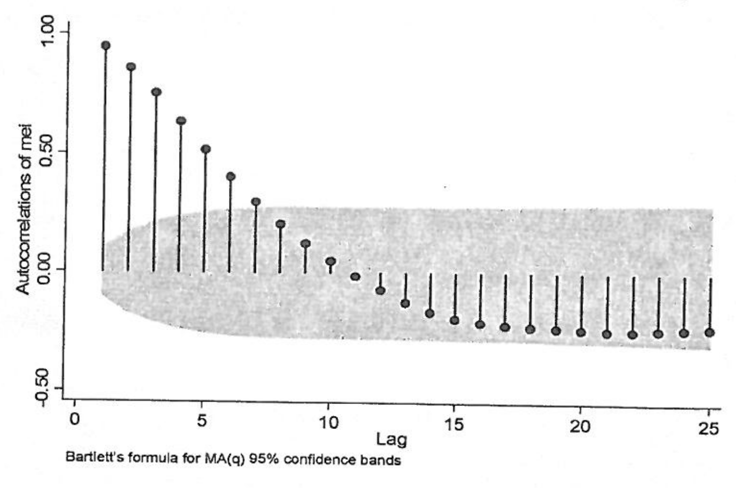
LAG	AC	PAC	Q	Prob>Q	-1 0 1 [Autocorrelation]	-1 0 1 [Partial Autocor]
1	0.9473	0.9569	336.5	0.0000		
2	0.8582	-0.4181	613.45	0.0000		
3	0.7532	-0.0631	827.33	0.0000		
4	0.6350	-0.1578	979.77	0.0000		
5	0.5167	0.0033	1081	0.0000		
6	0.4036	-0.0680	1142.9	0.0000		
7	0.2983	-0.0299	1176.8	0.0000		
8	0.2060	-0.0235	1193	0.0000		
9	0.1224	-0.0393	1198.8	0.0000		
10	0.0499	-0.0185	1199.7	0.0000		
11	-0.0140	-0.0359	1199.8	0.0000		
12	-0.0723	-0.0340	1201.8	0.0000		
13	-0.1243	-0.0456	1207.8	0.0000		

فترات التباطؤ تظهر في الجانب الأيسر من الجدول يليها قيم الارتباط الذاتي (AC) والارتباط الذاتي الجزئي (PAC)، فمثلاً الارتباط بين mei_t و mei_{t-2} هو 0.8582 بينما الارتباط الذاتي الجزئي (المعدل لفترة تباطؤ 1) هو 0.4181−، إحصائيات *Q* (إحصائية Ljung-Box) تختبر سلسلة من فرضيات العدم التي بها الارتباطات الذاتية تتضمن فترة تباطؤ تساوي صفرًا، وحيث إن قيم الاحتمال *p*-values التي نراها هنا كلها صغيرة جداً، فيجب علينا رفض فرضيات العدم، ونستنتج من ذلك بأن مؤشر ENSO التراكمي (*mei*) يوضح ارتباطاً ذاتياً ذا معنوية، وإذا وجدنا بأنه لا توجد أي قيمة احتمال لإحصائية *Q* أقل 0.05 فإنه يمكننا الاستنتاج بأن السلسلة كانت ذات ضجة بيضاء white noise مع ارتباط ذاتي ليس ذا معنوية.

في الجانب الأيمن لمخرجات جدول الأمر **corrgram** حيث إن هذا الجدول تميّز بخطوط بيانية توضح الارتباطات الذاتية، والارتباطات الذاتية الجزئية، فحص مثل هذه الخطوط البيانية يلعب دوراً في اختيار نماذج السلاسل الزمنية، فالتمثيل البياني للارتباط الذاتي يمكن الحصول عليه من خلال الأمر **ac**:

```
.ac mei, lags(25)
```

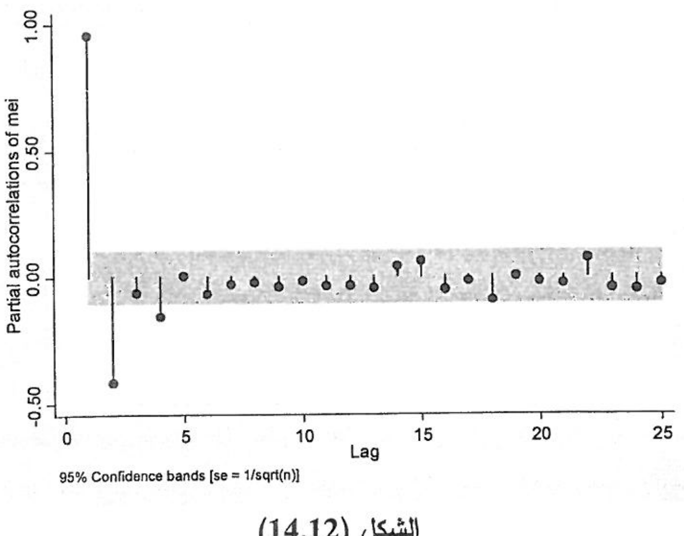
الشكل (13.12) يتضمن منطقة مظللة مع نقطة تمركز مع فترة ثقة 95%، الارتباطات خارج منطقة الثقة ذات معنوية إحصائية، الخيار **lags(25)** يوسّع هذا الشكل البياني إلى فترة تباطؤ 25 شهرًا موضحاً أن الارتباطات الذاتية للمتغير *mei* أصبحت سالبة بعد حوالي 12 شهرًا، هذا النمط يعكس خاصية شبه متذبذبة لـ ENSO: حيث إن الفترات العليا تميل لتكون متبوعة بفترات أقل وهكذا.



الشكل (13.12)

وبالمثل، فإن الأمر **pac** يقوم بإنشاء رسم بياني للارتباطات الذاتية الجزئية في الشكل (14.12). فترات الثقة التقريبية تقوم بناءً على تقديرات الخطأ المعياري لـ $\frac{1}{\sqrt{n}}$، الارتباطات الذاتية الجزئية للمتغير *mei* تختلف عن الارتباطات الذاتية، حيث إنها تقطع أغلب القيم التي ليست ذات معنوية بعد فترة تباطؤ تساوي شهرين.

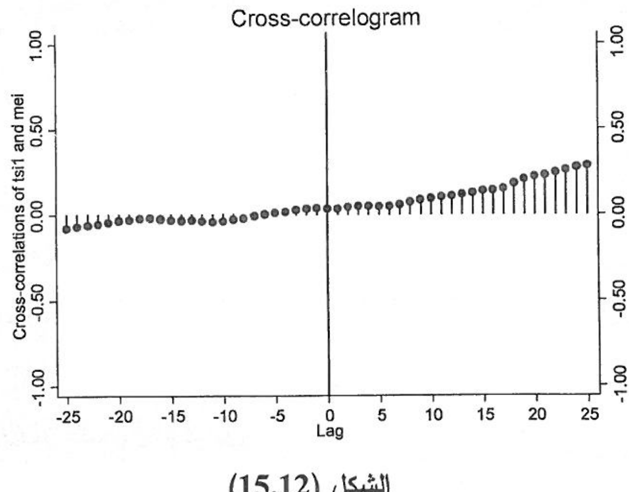
```
.pac mei, lags(25)
```



الشكل (14.12)

التمثيل البياني المتقاطع للارتباط الذاتي يساعد في اكتشاف العلاقة بين اثنين من السلاسل الزمنية، الشكل (15.12) يعرض شكلاً بيانياً متقاطعاً للمجموع الكلي لأشعة الشمس (*tsil*) و*mei*، التمثيل البياني المتقاطع للارتباط الذاتي يقترب من الصفر لفترات التباطؤ السالبة ثم يصبح أكثر قوة تدريجياً (بالرغم من أنه مازال ضعيفاً) عند فترات تباطؤ موجبة وأكثر طولاً.

```
.xcorr tsi1 mei, lags(25) xline(0) xlabel(-
 25(5)25)
```



الشكل (15.12)

التمثيل البياني المتقاطع للارتباط الذاتي هو طريقة سهلة جداً للقراءة إذا قمنا بوضع قائمة للمدخلات أو أول متغير مستقل في الأمر **xcorr** والمخرجات أو ثاني متغير تابع كما تم القيام به في الشكل (15.12)، يليها فترات تباطؤ موجبة تُشير إلى ارتباطات بين مدخلات عند الزمن *t* ومخرجات عند الزمن *t*+1، *t*+2 وهكذا. في هذا الشكل نرى بعض الارتباط بين أشعة الشمس وحالة ENSO حول سنتين لاحقاً ومعتدلاً بثبات، وهذا يؤدي إلى تأخير تأثير الشمس على ENSO، ليس هناك ارتباط بين أشعة الشمس وحالة ENSO خلال سنتين سابقتين ولا يعطي (بدرجة معقولة) أي سبب للتفكير بأن ENSO يؤثر على الشمس.

مُعَامِلات الارتباط المتقاطع، والإصدار النصي للتمثيل البياني للارتباط المتقاطع يمكن الحصول عليه باستخدام الخيار **table**:

```
.xcorr tsi1 mei, lags(13) table
```

```
                           -1       0       1
 LAG      CORR      [Cross-correlation]
-----------------------------------------------
-13      -0.0308                |
-12      -0.0370                |
-11      -0.0434                |
-10      -0.0372                |
-9       -0.0275                |
-8       -0.0209                |
-7       -0.0062                |
-6        0.0038                |
-5        0.0093                |
-4        0.0156                |
-3        0.0262                |
-2        0.0343                |
-1        0.0355                |
0         0.0326                |
1         0.0340                |
2         0.0427                |
3         0.0466                |
4         0.0468                |
5         0.0441                |
6         0.0476                |
7         0.0558                |
8         0.0721                |
9         0.0849                |
10        0.0930                |
11        0.1008                |
12        0.1041                |
13        0.1142                |
```

بقية هذا الفصل تتحدث عما بعد الارتباط لتتناول نماذج زمنية أكثر.

نماذج (ARIMA) : ARIMA Models

نماذج المتوسط المتحرك للانحدار الذاتي التكاملي Autoregressive integrated moving average (ARIMA) يمكن تقديرها من خلال الأمر **arima**، هذا الأمر يتضمن الانحدار الذاتي (AR) أو المتوسط المتحرك (MA) أو نماذج ARIMA، كما يمكن تقدير النماذج الهيكلية التي تتضمن متغيراً تنبؤياً أو أكثر، كما تتضمن أخطاء ARIMA، وهي تسمى نماذج ARMAX وتكون في صيغة مصفوفة كما يلي:

$$y_t = \mathbf{x}_t\beta + \mu_t \qquad [1.12]$$

حيث إن y_t مُوجّه قيم المتغير التابع عند الزمن t، $\mathbf{x}_t$ مصفوفة قيم المتغيرات التنبؤية الخارجية (وتتضمن في العادة ثابت) وμ_t هو موجّه الأخطاء "كل الأشياء الأخرى"، هذه الأخطاء يمكن أن تكون انحدارًا ذاتياً أو متوسطاً متحركاً من أي درجة، فمثلاً أخطاء ARMA(1,1) هي:

$$\mu_t = \rho\mu_{t-1} + \theta \in_{t-1} + \in_t \qquad [2.12]$$

حيث إن ρ معلمية الانحدار الذاتي من الدرجة الأولى، θ معلمية المتوسط المتحرك من الدرجة الأولى، $\in_t$ تمثل أخطاء الضجة البيضاء white noise وهي غير مترابطة وعشوائية، **arima** تناسب النماذج البسيطة مثل الحالة الخاصة في المعادلة [1.12] و[2.12] مع ثابت (β_0) يقوم باستبدال الصيغة الهيكلية $\mathbf{x}_t\beta$ وبالتالي فإن نموذج ARMA(1,1) البسيط يصبح:

$$\begin{aligned} y_t &= \beta_0 + \mu_t \\ &= \beta_0 = \rho\mu_{t-1} + \theta \in_{t-1} + \in_t \end{aligned} \qquad [3.12]$$

بعض المصادر تعرض أشكالاً بديلة، ففي حالة ARMA(1,1) فإنها تعرض y_t كدالة لقيمة y السابقة (y_{t-1}) وتعرض ($\in_t$) وأخطاء تباطؤية ($\in_{t-1}$):

$$y_1 = \alpha + \rho y_{t-1} + \theta \in_{t-1} + \in_t \qquad [4.12]$$

وبما أن النموذج الهيكلي المبسط يحتوي $y_t = \beta_0 + \mu_t$، فإن المعادلة [3.12] (المستخدمة من قبل برنامج ستاتا) تعادل المعادلة [4.12] كجزء من إعادة قياس الثابت $\alpha = (1-\rho)\beta_0$.

وباستخدام الأمر **arima**، فإن نموذج ARMA(1,1) يمكن تبسيطه بإحدى الطريقتين التاليتين:

```
.arima y, ar(1) ma(1)
```

أو:

```
.arima y, arima(1,0,1)
```

الحرف **i** في الأمر **arima** هو اختصار لكلمة "متكامل" "integrated" والتي تشير إلى النماذج التي تتضمن اختلافات، ولصياغة نموذج ARIMA(2,1,1) نقوم باستخدام:

```
.arima y, arima(2,1,1)
```

وهذا يكافئ:

```
.arima D.y, ar(1/2) ma(1)
```

لاحظ بأن الخيار **arima()** هو الرقم المُعطى لصيغة الانحدار الذاتي أو المتوسط المتحرك الذي يحدد كل فترات التباطؤ، والتي تتضمن ذلك الرقم، إذن فإن الرقم "2" يعني فترتي تباطؤ 1 و2. وعند استخدام الخيارات **ma()** أو **ar()** فإن العدد المستخدم يقوم بتحديد فترة تباطؤ معينة فقط. إذن فإن الرقم "2" يعني فترة تباطؤ تساوي 2 فقط. الأمران أعلاه يحددان نموذجاً به الاختلافات الأولى للمتغير التابع $(y_t - y_{t-1})$ هو دالة للاختلافات الأولى مع فترة تباطؤ واحدة وفترتي تباطؤ سابقة $(y_{t-1} - y_{t-2}$ و $y_{t-2} - y_{t-3})$ وكذلك للأخطاء الحالية والسابقة $(\in_t$ و $\in_{t-1})$.

لتقدير نموذج هيكلي والذي به y_t تعتمد على متغيرين تنبؤيين x (قيم حالية وتباطؤية x_t وx_{t-1}) وw (القيم الحالية فقط w_t) مع أخطاء ARIMA(1,0,3) وأيضاً الأخطاء ARIMA(1,0,1) الموسمية المضاعفة والتي تعود إلى 12 فترة زمنية (حيث إنها تكون مناسبة للبيانات الشهرية خلال عدد من السنوات) والأمر المناسب للقيام بذلك قد يأخذ الصيغة التالية:

```
.arima y x L.x w, arima(1,0,3) sarima(1,0,1,12)
```

وفي سياق الاقتصاد القياسي، فإن هذا يتوافق مع نموذج $ARIMA(1,0,3)\times(1,0,1)_{12}$

السلسلة الزمنية y تُعتبر مستقرة إذا كان متوسطها وتباينها لم يتغيرا مع الوقت، وإذا كان التغاير المصاحب بين y_t و$y_{t+\mu}$ يعتمد فقط على فترة تباطؤ μ

وليس على قيم معينة لـــ t، صياغة نماذج ARIMA تفتـــرض أن السلـــسلة الزمنية مستقرة أو يمكن جعلها مستقرة من خلال طرق مناسبة للتحويـــل أو الفروقات. يمكننا فحص هذه الفرضية بواسطة فحص الشكل البياني لاتجـــاه الزمن من خلال مستوى التباين. رأينا سابقاً أن درجات الحرارة العالمية مثلاً شهدت اتجاهاً معيناً يوضح بأنها غير مستقرة.

الاختبارات الإحصائية الرسمية لـــ "جذور الوحدة" غير مستقرة في طريقـــة AR(1) والتي تكون فيها $\rho_1 = 1$ (والتي تُعرف أيضاً بأنها المسار العشوائي) تعتبر نظرية المتممة و فحص غير رسمي، برنامج ستاتا يوفر ثلاثة اختبارات لجـــذر الوحدة هي: **pperron** (فيليبس– برون Phillips-Perron)، **dfuller** (ديكي – فـــولر المتمم augmented Dickey-Fuller)، **dfgls** (ديكي – فولر المتمم باســـتخدام GLS)؛ ويُعتبر اختبار **dfgls** أقوى الاختبارات وأكثر معلومات.

الجدول أدناه يُطبق الأمر **dfgls** على بيانات المركز الـــوطني للمنـــاخ (NCDC) وهي درجات الحرارة العالمية الشاذة.

```
.dfgls ncdctemp, notrend
DF-GLS for ncdctemp                                  Number of obs =   355
Maxlag = 16 chosen by Schwert criterion

              DF-GLS mu       1% Critical        5% Critical      10% Critical
 [lags]     Test Statistic        Value             Value             Value
------------------------------------------------------------------------------
   16          -1.162           -2.580            -1.952            -1.637
   15          -1.191           -2.580            -1.955            -1.640
   14          -1.179           -2.580            -1.958            -1.643
   13          -1.170           -2.580            -1.962            -1.646
   12          -1.194           -2.580            -1.965            -1.649
   11          -1.264           -2.580            -1.968            -1.652
   10          -1.190           -2.580            -1.971            -1.655
   9           -1.204           -2.580            -1.974            -1.658
   8           -1.525           -2.580            -1.977            -1.660
   7           -1.462           -2.580            -1.980            -1.663
   6           -1.639           -2.580            -1.982            -1.665
   5           -1.844           -2.580            -1.985            -1.668
   4           -1.968           -2.580            -1.988            -1.670
   3           -2.058           -2.580            -1.990            -1.672
   2           -2.342           -2.580            -1.993            -1.675
   1           -2.731           -2.580            -1.995            -1.677

Opt Lag (Ng-Perron seq t) =  9 with RMSE  .0915337
Min SC   = -4.687887 at lag  1 with RMSE  .0943745
Min MAIC = -4.721196 at lag  9 with RMSE  .0915337
```

مخرجات الأمر **dfgls** أعلاه توضح بأن اختبارات عدم استقرارية فرضية العدم التي تقول بأن سلسلة درجات الحرارة تمثل مسارًا عشوائياً أو لها جذر وحدة لفترات تباطؤ من 1 إلى 16 شهرًا، وفي أسفل جدول المخرجات هناك ثلاث طرق مختلفة لاختيار العدد المناسب من فترات التباطؤ: Ng- بيرون التتابعي *t*Ng-Perron sequential، معيار المعلومات شوارتز Schwarz الأقل، Ng- بيرون لمعيار معلومات أكيكا Akaike المعدلة (MAIC)، MAIC والذي تم تطويره مؤخراً، وتجربة مونت كارلو Monte Carlo كلاهما يدعم مميزات طريقة شوارتز. إن انتظام *t* وMAIC توصي باستخدام 9 فترات تباطؤ. كما أن النتائج توضح بأن إحصائية DF-GLS لـ 9 فترات تباطؤ هي 1.204- وهي أكبر من 10% من القيمة الحرجة وهي 1.658- ولذلك يجب علينا رفض فرضية العدم. هذه النتائج تؤكد الانطباعات السابقة، والتي تقول بأن السلسلة الزمنية *ncdctemp* ليست مستقرة.

ومن ناحية أخرى، فإن هناك اختباراً مشابهاً لمؤشر ENSO المتعدد يرفض فرضية عدم الاستقرار عند كل فترات التباطؤ حتى مستوى 1%.

```
.dfgls mei, notrend
DF-GLS for mei                                        Number of obs =   355
Maxlag = 16 chosen by Schwert criterion

           DF-GLS mu        1% Critical      5% Critical      10% Critical
 [lags]  Test Statistic        Value            Value             Value
-----------------------------------------------------------------------------
   16       -3.686            -2.580           -1.952           -1.637
   15       -3.742            -2.580           -1.955           -1.640
   14       -3.681            -2.580           -1.958           -1.643
   13       -4.062            -2.580           -1.962           -1.646
   12       -4.381            -2.580           -1.965           -1.649
   11       -4.352            -2.580           -1.968           -1.652
   10       -4.420            -2.580           -1.971           -1.655
   9        -4.451            -2.580           -1.974           -1.658
   8        -4.589            -2.580           -1.977           -1.660
   7        -4.604            -2.580           -1.980           -1.663
   6        -4.655            -2.580           -1.982           -1.665
   5        -4.699            -2.580           -1.985           -1.668
   4        -4.591            -2.580           -1.988           -1.670
   3        -4.909            -2.580           -1.990           -1.672
   2        -4.293            -2.580           -1.993           -1.675
   1        -4.049            -2.580           -1.995           -1.677

Opt Lag (Ng-Perron seq t) =  3 with RMSE  .2688427
Min SC   = -2.565539 at lag  1 with RMSE  .2727197
Min MAIC = -2.497381 at lag  1 with RMSE  .2727197
```

بالنسبة للسلاسل المستقرة مثل *mei*، فإن التمثيل البياني للارتباط يعتبر دليلاً حول اختيار نموذج ARIMA المبدئي:

AR(p) الانحدار الذاتي من الدرجة p، وله ارتباطات ذاتية تنخفض بشكل تدريجي مع كل زيادة في فترات التباطؤ، الارتباطات الذاتية الجزئية تتوقف بعد فترة تباطؤ p.

MA(q) المتوسط المتحرك من الدرجة q، وله ارتباطات ذاتية تتوقف بعد فترة تباطؤ q، الارتباطات الذاتية الجزئية تنخفض تدريجياً مع كل زيادة في فترات التباطؤ.

ARMA(p,q) المتوسط المتحرك – الانحدار الذاتي المختلط والذي له ارتباطات ذاتية، وارتباطات ذاتية جزئية تنخفض تدريجياً مع كل زيادة في فترات التباطؤ.

المنحنى البياني للارتباط يزداد بدرجة كبيرة عند فترات تباطؤ موسمية (مثلاً عند 12 أو 24 أو 36 في البيانات الشهرية) مشيراً إلى نمط موسمي، تحديد النماذج الشهرية يتبع نفس الطريقة التي تم تطبيقها على الارتباطات الذاتية، والارتباطات الذاتية الجزئية عند فترات تباطؤ موسمية.

رأينا سابقاً أن الارتباطات الذاتية للمتغير *mei* تنخفض تدريجياً مع زيادة فترات التباطؤ (الشكل 14.12) بينما الارتباطات الذاتية الجزئية تتوقف بعد فترة تباطؤ تساوي 2. هذا النمط في التمثيل البياني للارتباط مع نتائج اختبار **dfgls** تدعم الاستقرار، وتشير إلى أنه بالإمكان صياغة نموذج للمتغير *mei* وذلك بعملية ARIMA(2,0,0).

```
.arima mei, arima(2,0,0) nolog
ARIMA regression

Sample:  Jan1980 - Dec2010                      Number of obs      =       372
                                                Wald chi2(2)       =   4385.66
Log likelihood = -43.61494                      Prob > chi2        =    0.0000

------------------------------------------------------------------------------
             |                 OPG
         mei |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
mei          |
       _cons |   .2814275   .2132974     1.32   0.187    -.1366278    .6994828
-------------+----------------------------------------------------------------
ARMA         |
          ar |
         L1. |   1.349918   .0424774    31.78   0.000     1.266664    1.433173
         L2. |  -.4163392   .0415425   -10.02   0.000     -.497761   -.3349174
-------------+----------------------------------------------------------------
      /sigma |   .2710638   .0091183    29.73   0.000     .2531922    .2889354
------------------------------------------------------------------------------
Note: The test of the variance against zero is one sided, and the two-sided confidence interval is
      truncated at zero.
```

نموذج ARIMA(2,0,0) يمثل المتغير *mei* كدالة للأخطاء (μ) من شهر ماضي وشهرين ماضيين، بالإضافة إلى أخطاء الضجة البيضاء العشوائية ($\in$):

$$y_t = \beta_0 + \rho_t\mu_{t-1} + \rho_2\mu_{t-2} + \in_t \qquad [5.12]$$

حيث إن: y_t قيمة المتغير *mei* عند الزمن *t*، مخرجات الجدول توضح معلمية لتقدير $\beta_0 = 0.28$ ، $\rho_1 = 1.35$ ، $\rho_1 = 0.42$.

بعد صياغة نموذج **arima** هذه المُقدّرات والنتائج الأخرى يتم حفظها بشكل مؤقت بطريقة ستاتا المعتادة، فمثلاً لمشاهدة معامل النموذج AR(1) والخطأ المعياري نقوم بطباعة الأمر:

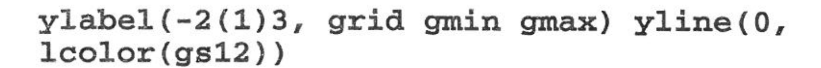
```
.display [ARMA]_b[L1.ar]
 1.3499184
.display [ARMA]_se[L1.ar]
 .04247738
```

معاملات الانحدار الذاتي من الدرجة الأولى والدرجة الثانية في هذا المثال، كلها ذات معنوية إحصائية وبعيدة جداً عن الصفر، (*t*=31.78 و-10.02 على التوالي) وهذا يشير إلى شيء واحد، وهو أن النموذج مناسب، ويمكننا الحصول على القيم المتوقعة والبواقي والإحصائيات الأخرى بعد الأمر **arima** من خلال الأمر **predict**:

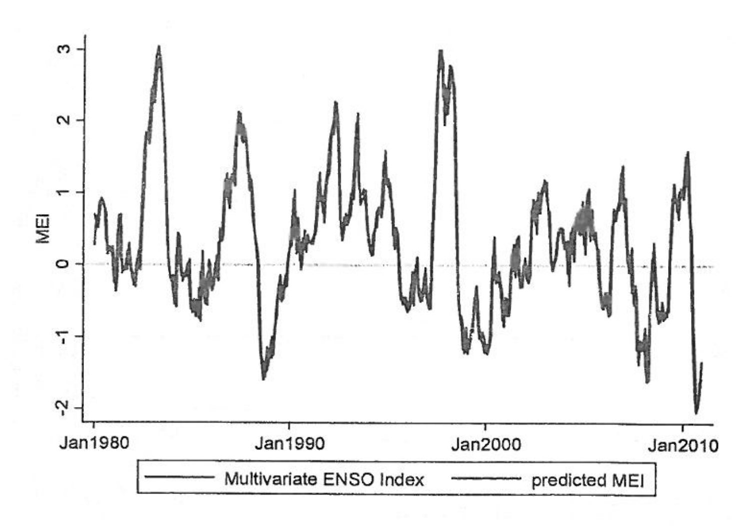
```
.predict meihat
.label variable meihat "predicted MEI"
.predict meires, resid
.label variable meires "residual MEI"
```

بيانياً، فإن القيم المتوقعة من هذا النموذج تظهرغير واضحة من درجات الحرارة المشاهدة (الشكل 16.12). هذه الصورة توضح مدى تقارب نماذج ARIMA، والتي يمكنها أن تتناسب بدرجة كبيرة مع السلاسل المرتبطة ذاتياً من خلال التنبؤ بمتغير من خلال قيمه السابقة، بالإضافة إلى قيم الأخطاء السابقة، نموذج ARIMA(2,0,0) يُفسّر حوالي 90% من تباين المتغير *mei*.

```
.tsline meimeihat, lcolor(blue red)
 lwidth(medium medthick)
 xtitle("") xlabel(, grid gmax gmin)
 ytitle("MEI")
```



```
ylabel(-2(1)3, grid gmin gmax) yline(0,
lcolor(gs12))
```



الشكل (16.12)

اختبار مهم لكفاءة نموذج ARIMA يتم من خلال ما إذا كانت البواقي تظهر غير مترابطة مع التشويش. الأمر **corrgram** يختبر فرضية العدم للضجة البيضاء من خلال فترات تباطؤ مختلفة.

```
.corrgram meires, lags(13)

                                          -1       0       1 -1       0       1
 LAG       AC       PAC      Q     Prob>Q  [Autocorrelation]  [Partial Autocor]
-------------------------------------------------------------------------------
 1      -0.0241  -0.0241   .2176   0.6409          |                  |
 2      -0.0099  -0.0105  .25451   0.8805          |                  |
 3       0.1430   0.1430  7.9641   0.0468          |-                 |-
 4      -0.0015   0.0049   7.965   0.0929          |                  |
 5       0.0256   0.0292  8.2135   0.1449          |                  |
 6       0.0200   0.0010  8.3656   0.2125          |                  |
 7      -0.0305  -0.0315  8.7197   0.2734          |                  |
 8      -0.0001  -0.0109  8.7197   0.3665          |                  |
 9      -0.0306  -0.0367  9.0787   0.4300          |                  |
 10     -0.0274  -0.0256  9.3682   0.4976          |                  |
 11     -0.0292  -0.0331  9.6962   0.5579          |                  |
 12     -0.0009   0.0079  9.6966   0.6426          |                  |
 13     -0.0449  -0.0404  10.477   0.6545          |                  |
```

اختبارات ***Q*corrgram** portmanteau لم تجد أي ارتباط ذاتي ذي معنوية إحصائية بين البواقي حتى فترة تباطؤ تساوي 13، باستثناء علاقة واحدة فقط

بمستوى أقل من 0.05 عند فترة التباطؤ 3، يمكننا القيام بنفس الاختبار لأي فترة تباطؤ معنية باستخدام الأمر **wntestq** مثل:

```
.wntestq meires, lags(13)

 Portmanteau test for white noise
 ---------------------------------------
  Portmanteau (Q) statistic =   10.4772
  Prob > chi2(13)           =    0.6545
```

ولذا، فإن نموذج ARIMA(2,0,0) تم اقتراحه بعد النظر إلى الأنماط الموجودة ببيانات الارتباط الذاتي، والتي توضح بأن النموذج يتناسب بدرجة كبيرة، حيث إن كل صيغ الخطأ AR ذات معنوية إحصائية، وهذا يترك البواقي التي اجتازت اختبار الضجة البيضاء.

اختباراتنا السابقة **dfgls**، **notrend** لدرجات الحرارة العالمية الشاذة (*ncdctemp*)– تختلف عن اختبار مؤشر ENSO التراكمي (*mei*)– حيث إنها توضح بأن درجات الحرارة تظهر غير مستقرة بسبب أنها تتضمن اتجاهاً معيناً مسيطراً، اختبار **dfgls** الافتراضي يتضمن اتجاهاً خطياً يرفض فرضية العدم التي تفترض عدم الاستقرار عند كل فترات التباطؤ.

```
.dfgls ncdctemp

DF-GLS for ncdctemp                                    Number of obs =   355
Maxlag = 16 chosen by Schwert criterion

               DF-GLS tau      1% Critical       5% Critical      10% Critical
  [lags]     Test Statistic       Value             Value             Value
 ------------------------------------------------------------------------------
    16           -3.624          -3.480            -2.818            -2.536
    15           -3.625          -3.480            -2.824            -2.542
    14           -3.552          -3.480            -2.829            -2.547
    13           -3.489          -3.480            -2.835            -2.552
    12           -3.486          -3.480            -2.840            -2.557
    11           -3.566          -3.480            -2.846            -2.562
    10           -3.395          -3.480            -2.851            -2.566
    9            -3.377          -3.480            -2.856            -2.571
    8            -3.861          -3.480            -2.861            -2.575
    7            -3.697          -3.480            -2.865            -2.580
    6            -3.950          -3.480            -2.870            -2.584
    5            -4.251          -3.480            -2.874            -2.588
    4            -4.399          -3.480            -2.879            -2.592
    3            -4.479          -3.480            -2.883            -2.595
    2            -4.920          -3.480            -2.887            -2.599
    1            -5.505          -3.480            -2.891            -2.602

Opt Lag (Ng-Perron seq t) =  9 with RMSE  .0902188
Min SC   = -4.749446 at lag  1 with RMSE  .0915139
Min MAIC = -4.639686 at lag  9 with RMSE  .0902188
```

الاختلاف الأولي للسلاسل الزمنية، سوف يزيل الاتجاه الخطي، الاختلافات الأولية لدرجات الحرارة توضح نمطاً من الارتباطات الذاتية التي تتوقف بعد فترة تباطؤ تساوي 1، والارتباطات الذاتية الجزئية تنخفض بعد فتراة تباطؤ تساوي 1.

```
.corrgram D.ncdctemp, lag(13)

                                          -1       0       1 -1       0       1
 LAG       AC       PAC      Q     Prob>Q  [Autocorrelation]  [Partial Autocor]
-------------------------------------------------------------------------------
1       -0.3941  -0.4031   58.094  0.0000       ---|                ---|
2        0.0425  -0.1355   58.773  0.0000          |                  -|
3       -0.0639  -0.1174   60.307  0.0000          |                   |
4        0.0305  -0.0479   60.656  0.0000          |                   |
5       -0.0359  -0.0614   61.145  0.0000          |                   |
6       -0.0385  -0.1032   61.706  0.0000          |                   |
7       -0.0035  -0.0872   61.711  0.0000          |                   |
8        0.0633   0.0164   63.239  0.0000          |                   |
9       -0.1316  -0.1432   69.861  0.0000         -|                  -|
10       0.1031  -0.0149   73.938  0.0000          |                   |
11       0.0030   0.0359   73.942  0.0000          |                   |
12      -0.0353  -0.0393   74.421  0.0000          |                   |
13       0.0099  -0.0175   74.459  0.0000          |                   |
```

هذه المشاهدات تشير إلى أن نموذج ARIMA(0,1,1) قد يكون مناسباً لدرجات الحرارة.

```
.arima ncdctemp, arima(0,1,1) nolog
ARIMA regression

Sample:  Feb1980 - Dec2010                    Number of obs      =       371
                                              Wald chi2(1)       =    152.87
Log likelihood =  355.3385                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
             |                 OPG
 D.ncdctemp  |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ncdctemp     |
       _cons |   .0007623   .0025694     0.30   0.767    -.0042737    .0057982
-------------+----------------------------------------------------------------
ARMA         |
          ma |
         L1. |  -.4994929    .040399   -12.36   0.000    -.5786735   -.4203124
-------------+----------------------------------------------------------------
      /sigma |   .0928235   .0029514    31.45   0.000     .0870388    .0986081
------------------------------------------------------------------------------
```

Note: The test of the variance against zero is one sided, and the two-sided confidence interval is truncated at zero.

نموذج ARIMA(0,1,1) يشرح الاختلاف الأول أو التغير من شهر إلى شهر في درجة الحرارة كدالة للخطأ العشوائي مع فترة تباطؤ تساوي شهراً واحداً أو كدالة للخطأ العشوائي للشهر الحالي:

$$y_t - y_{t-1} = \beta_0 + \theta \in_{t-1} + \in_t \qquad [6.12]$$

حيث إن: y_t تمثل المتغير *ncdctemp* عند الزمن t، ومُقدِّرات المعلميات هي $\beta_0 = 0.00076$ و $\theta = -0.499$، حد MA من الدرجة الأولى θ ذو معنوية إحصائية ($p \approx 0.000$) وبواقي النموذج لا يمكن تمييزها عن الضجة البيضاء.

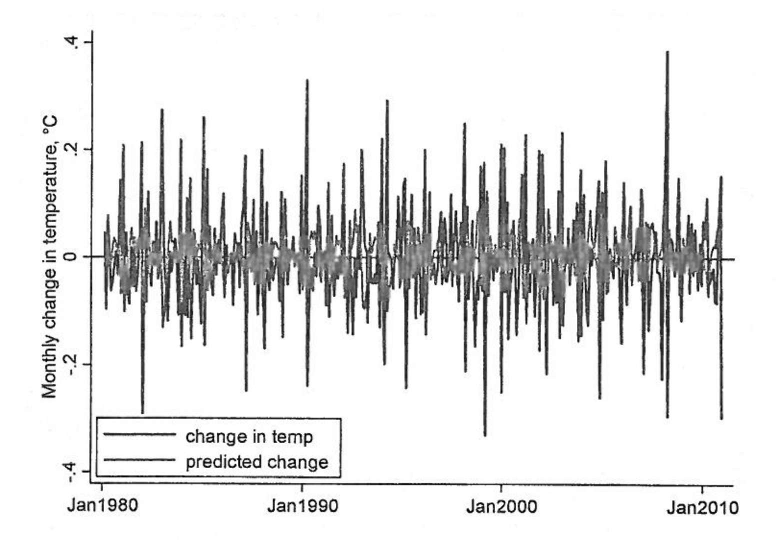
```
.predict Dncdchat
.label variable Dncdchat "predicted 1st diff temp"
.predict Dncdcres, resid
.label variable Dncdcres "residual 1st diff temp"
.corrgram Dncdcres, lag(13)
```

LAG	AC	PAC	Q	Prob>Q	-1 0 1 [Autocorrelation]	-1 0 1 [Partial Autocor]
1	0.0133	0.0135	.06574	0.7976		
2	0.0262	0.0265	.32335	0.8507		
3	-0.0631	-0.0657	1.8204	0.6105		
4	-0.0313	-0.0306	2.1907	0.7007		
5	-0.0848	-0.0827	4.912	0.4267		
6	-0.0867	-0.0910	7.7626	0.2560		
7	-0.0357	-0.0362	8.2477	0.3113		
8	0.0079	-0.0002	8.2716	0.4074		
9	-0.0970	-0.1201	11.866	0.2210		
10	0.0759	0.0635	14.074	0.1696		
11	0.0255	0.0143	14.324	0.2156		
12	-0.0225	-0.0568	14.519	0.2688		
13	0.0008	-0.0004	14.519	0.3383		

بالرغم من أن هذه الاختبارات لا تعطي سبباً واضحاً لاستبعاد ARIMA(0,1,1)، إلا أنه من الصعب تنبؤ التغيرات الشهرية في درجات الحرارة العالمية الشاذة، الشكل (17.12) يستخدم مُعامِل الاختلاف .D في

الأمر graph لعرض تناسب بسيط بين القيم المتوقعة والقيم المشاهدة، هـذا النموذج يشرح حوالي 20% من التباين في الاختلافات الشهرية.

```
.tsline D.ncdctempDncdchat, lcolor(blue red)
     xtitle("")  xlabel(,  grid  gmax  gmin)
     lw(medthick medium)
     ytitle("Monthly  change  in  temperature,
     `=char(176)'C")
     ylabel(,  grid  gmin  gmax)  yline(0)
     legend(ring(0) position(7)
     row(2) label(1 "change in temp") label(2
     "predicted change"))
```



الشكل (17.12)

الرسم البياني للتغيرات أو الاختلافات الأولـى فـي الـشكل (17.12) يعطي تشابهاً بسيطاً مع درجات الحرارة الشاذة نفسها التي سبق مشاهدتها في الشكل (9.12)، العكس يؤكد بأن صياغة نماذج للاختلافات الأولى يعطـي الإجابة عن سؤال بحث مختلف.

في هذا المثال الميزة الرئيسة لدرجات الحرارة المسجلة – الاتجاه نحـو الارتفاع – تم استبعاده؛ الجزء التالي يعود إلى درجات الحرارة الشاذة غيـر المختلفة ويأخذ في الاعتبار كيفية إمكانية شرح هذا الاتجاه نفسه.

نماذج (ARMAX) ARMAX Models :

سابقاً في هذا الفصل، رأينا أن انحدار OLS للمتغير *ncdctemp* مع أربعة متغيرات تنبؤية بأربع فترات تباطؤ يعطي تناسباً جيدًا مع قيم درجات الحرارة المشاهدة (الشكل 2.12)، بالإضافة إلى تقديرات معلمية معقولة، اختبار دربن واتسون Durbin-Watson أظهر بأن هناك ارتباطات ذاتية ذات معنوية بين البواقي وهذا يقوّض اختبارات *F* و*t*OLS. ولحل هذه المشكلة نقوم باستخدام نماذج ARMAX (المتوسط المتحرك للانحدار الذاتي مع متغيرات خارجية).

نموذج ARMAX مع متغيرات تنبؤية بفترات تباطؤ لشهر واحد واضطراب ARIMA(1,0,1) اتضح بأنه يتناسب بشكل جيد.

```
.arima ncdctemp L1.aod L1.tsi1 L1.mei
 L1.co2anom, arima(1,0,1) nolog
```

```
ARIMA regression

Sample:  Feb1980 - Dec2010                    Number of obs    =       371
                                              Wald chi2(6)     =    555.93
Log likelihood =  378.3487                    Prob > chi2      =    0.0000
```

ncdctemp	Coef.	OPG Std. Err.	z	P>|z|	[95% Conf.	Interval]
ncdctemp						
aod L1.	-1.228967	.3855346	-3.19	0.001	-1.984601	-.4733331
tsi1 L1.	.0609574	.0173356	3.52	0.000	.0269803	.0949345
mei L1.	.0533736	.0099622	5.36	0.000	.033848	.0728992
co2anom L1.	.0104806	.0008328	12.58	0.000	.0088483	.0121128
_cons	-82.84697	23.68097	-3.50	0.000	-129.2608	-36.43313
ARMA						
ar L1.	.7119696	.0703746	10.12	0.000	.5740378	.8499013
ma L1.	-.3229314	.0944706	-3.42	0.001	-.5080903	-.1377725
/sigma	.0872355	.0028313	30.81	0.000	.0816863	.0927847

Note: The test of the variance against zero is one sided, and the two-sided confidence interval is truncated at zero.

في هذا النموذج y_t أو المتغير *ncdctemp* عند الزمن t هـو دالـة لقـيم المتغيرات التنبؤية من x_1 وحتى x_4 مع فترة تباطؤ تساوي 1 (هذه المتغيرات من *aod* وحتى *co2anom*) واضطراب (μ_t):

$$y_t = \beta_0 + \beta_1 x_{1,t-1} + \beta_2 x_{2,t-1} + \beta_3 x_{3,t-1} + \beta_4 x_{4,t-1} + \mu_t \qquad [7.12]$$

الاضطراب عند الزمن $t(\mu_t)$ يتعلق باضطراب مع فترة تباطؤ تساوي 1 (μ_{t-1})، ومتعلق كذلك بأخطاء عشوائية لها فترة تباطؤ تساوي 1 عند الزمن t ($\in_t$ و $\in_{t-1}$):

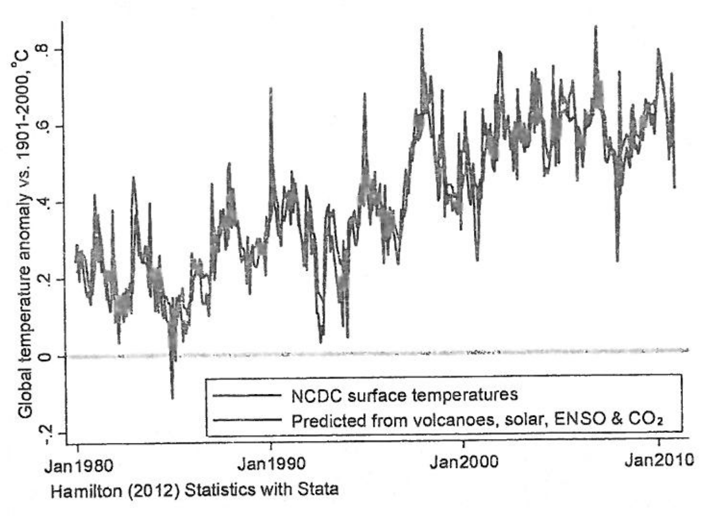
$$u_t - \rho\mu_{t-1} = \theta \in_{t-1} + \in_t \qquad [8.12]$$

مُعامِلات كل المتغيرات التنبؤية الأربعـة، وكـذلك حـدود AR وMA جميعها ذات معنوية إحصائية عند $p < 0.01$ أو أفضل مـن ذلـك، المتغيـر *ncdctemp* نفسه غير مستقر، وليست هناك حاجة ليكـون كـذلك فـي هـذا التحليل، وعدم استقراره هو نقطة التركيز في البحث. البواقي من أي نموذج ناجح يُفترض أن تتشابه مع قيمـة الضجة البيضاء، وقيمة استقرار التغاير، وهذا هو الوضع هنا. حيث إن: اختبـار Q يعطـي $p = 0.60$ للبواقي التـي لها فترات تباطؤ تصل حتى 25.

```
.predict ncdchat2
.label variable ncdchat2 "Predicted from
 volcanoes, solar, ENSO & CO2"
.predict ncdcres2, resid
.label variable ncdcres2 "NCDC residuals from
 ARMAX model"
.corrgram ncdcres2, lags(25)
```

LAG	AC	PAC	Q	Prob>Q	-1 0 1 [Autocorrelation]	-1 0 1 [Partial Autocor]
1	-0.0133	-0.0136	.06661	0.7963		
2	0.0458	0.0471	.85498	0.6521		
3	-0.0089	-0.0084	.88473	0.8291		
4	0.0248	0.0229	1.1169	0.8916		
5	-0.0359	-0.0356	1.6043	0.9007		
6	-0.0658	-0.0710	3.245	0.7775		
7	-0.0060	-0.0035	3.2588	0.8601		
8	0.0116	0.0177	3.3097	0.9134		
9	-0.0949	-0.0979	6.7554	0.6626		
10	0.0832	0.0868	9.4103	0.4937		
11	0.0341	0.0433	9.8585	0.5432		
12	0.0052	-0.0105	9.869	0.6275		
13	0.0388	0.0457	10.45	0.6568		
14	0.0330	0.0288	10.873	0.6960		
15	-0.0110	-0.0273	10.92	0.7583		
16	-0.0381	-0.0288	11.485	0.7786		
17	-0.0871	-0.0861	14.449	0.6351		
18	-0.0384	-0.0537	15.028	0.6601		
19	-0.0180	0.0130	15.155	0.7127		
20	-0.0508	-0.0484	16.174	0.7058		
21	-0.0619	-0.0764	17.688	0.6686		
22	0.0339	0.0422	18.143	0.6975		
23	0.0557	0.0551	19.379	0.6790		
24	0.0779	0.0652	21.802	0.5911		
25	0.0457	0.0457	22.638	0.5987		

الشكل (18.12) يعرض التناسب الكبير بين البيانات ونموذج ARMAX. النموذج يفسّر حوالي 77% من التباين في درجات الحرارة الشاذة. نتائج ARMAX تدعم نتيجة عامة من نتائج OLS السابقة، حيث إن الاتجاه المتصاعد خلال عقود متعددة في درجات الحرارة لا يمكن تفسيره بدون الأخذ في الاعتبار العوامل البشرية. ومن ناحية أخرى، فإن الانخفاض الواضح في الاحتباس الحراري خلال العقد الأخير يمكن تفسيره في ظل ظاهرة طبيعية إلينوي، وظاهرة انخفاض أشعة الشمس.



الشكل (18.12)

حتى هذه النقطة، لدينا تفاصيل عن مؤشر واحد لدرجة الحرارة من ثلاثة مؤشرات بملف البيانات *Climate.dta*. هذا المؤشر تم استخراجه من بيانات المركز الوطني للمناخ (NCDC) التابع للإدارة الوطنية للغلاف الجوي والمحيط بالولايات المتحدة (NOAA)، NCDC حيث قام بحساب مؤشره الخاص به بناءً على قياسات درجة حرارة السطح، والتي تم أخذها من آلاف المحطات حول العالم. مركز ناسا لدراسات الفضاء قام بحساب مؤشره الخاص بدرجة الحرارة (يُطلق عليه GISTEMP) والذي يعتمد على قياسات المحطات لدرجة حرارة السطح، ولكن مع تغطية جيدة للمناطق التي تقع في أقصى الشمال.

مؤشر درجة الحرارة الثالث في الملف *Climate.dta* له قاعدة مختلفة، فالباحثون بجامعة الباما هندسفيل (UAH) قاموا بحساب مؤشرهم العالمي من مقاييس غير مباشرة من الأقمار الصناعة التي قامت بتصوير طبقات الجو العليا وأجواء الأرض عند ارتفاع حوالي 4 كيلومترات، التقديرات التي تعتمد على الأقمار الصناعية توضح تغيرًا كبيرًا من شهر إلى شهر وهي تغيرات حساسة لظهور إلنينو ولانينا. هذه التقديرات تعتبر نظرة بديلة عن اتجاهات

درجات الحرارة العالمية، وهي تتشابه مع سجلات السطح، ولكنها تختلف قليلاً في التفاصيل. السؤال هنا: كيف يتم عرض هذه الاختلافات عندما نقوم بصياغة نموذج للمتغير *uahtemp* بنفس طريقة ARMAX التي تم استخدامها مع المتغير *ncdctemp*؟

```
.arima uahtemp L1.aod L1.tsi1 L1.mei L1.co2anom,arima(1,0,1) nolog
ARIMA regression

Sample:  Feb1980 - Dec2010                    Number of obs    =       371
                                              Wald chi2(6)     =    601.88
Log likelihood =  299.2819                    Prob > chi2      =    0.0000
```

uahtemp	Coef.	OPG Std. Err.	z	P>|z|	[95% Conf.	Interval]
uahtemp						
aod L1.	-2.38566	.9011263	-2.65	0.008	-4.151835	-.6194849
tsi1 L1.	.0336446	.0289365	1.16	0.245	-.0230698	.0903591
mei L1.	.0663992	.0154607	4.29	0.000	.0360967	.0967016
co2anom L1.	.0084778	.0016671	5.09	0.000	.0052103	.0117453
_cons	-45.92206	39.52334	-1.16	0.245	-123.3864	31.54227
ARMA						
ar L1.	.8364133	.0421928	19.82	0.000	.7537169	.9191097
ma L1.	-.3170064	.068849	-4.60	0.000	-.451948	-.1820648
/sigma	.1078988	.0040726	26.49	0.000	.0999167	.115881

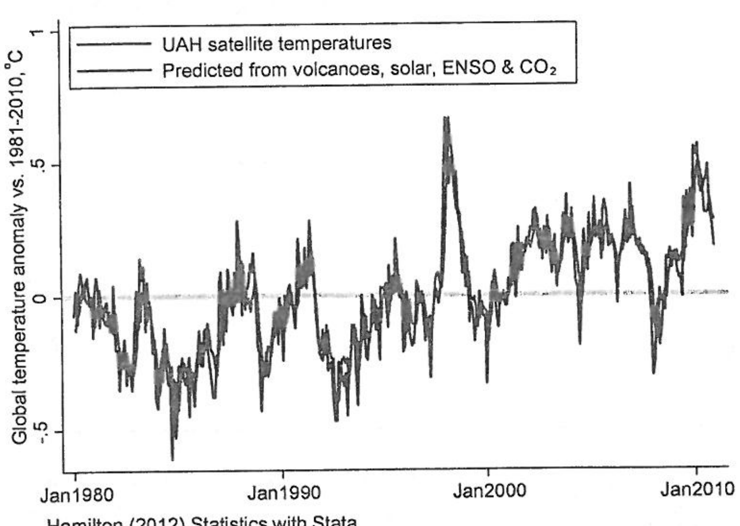
Note: The test of the variance against zero is one sided, and the two-sided confidence interval is truncated at zero.

معاملات نموذج NCDC ونموذج UAH– كما أن معاملات NASA لم يتم عرضها هنا – كلها لها نفس الإشارة وتتشابه مع الفهم الطبيعي. كل النماذج الثلاثة تؤكد بأن أقوى متغير تنبؤي لدرجات الحرارة هو القيم الشاذة لـــ CO_2

وثاني أقوى متغير تنبؤي هو MEI. المتغير MEI له تأثير أكبر نسبياً في نموذج UAH، وهذا هو المتوقع من بيانات الأقمار الصناعية، كما أن ثورات البراكين لها أيضاً تأثيرات قوية على مؤشر UAH، بينما أشعة الشمس كانت لها تأثيرات أقل قوة، هذا يعكس التغير المرتفع في التقديرات التي تعتمد على الأقمار الصناعة، كما أن النموذج يشرح حوالي 70% من التباين في المتغير *uahtemp* مقارنة مع 77% للمتغير *ncdctemp*، وعموماً فإن البواقي اجتازت الاختبارات الخاصة بالضجة البيضاء ($p = 0.65$)، والرسم البياني للقيم المتوقعة والمشاهدة يعرض تناسباً مرئياً جيدًا (الشكل 19.12).

```
.predict uahhat2
.label variable uahhat2 "Predicted from volcanoes,
solar, ENSO & CO2"
.predict uahres2, resid
.label variable uahres2 "UAH residuals from ARMAX model"
.wntestq uahres2, lags(25)

Portmanteau test for white noise
---------------------------------------
Portmanteau (Q) statistic =    21.7197
Prob > chi2(25)           =     0.6519
```



الشكل (19.12)

درجات الحرارة المسجلة بالأقمار الصناعية تمثل بيانات أكثر حداثة من سجلات محطات الإرصاد الجوية، ولذلك فإن درجات الحرارة الشاذة UAH يمكن حسابها للفترة من 1981–2010 كنقطة بداية بدلاً من 1901–2000. كما تم القيام به من قبل NCDC. نقطة البداية تنتقل إلى نقطة الصفر لدرجات الحرارة الشاذة الأعلى، والذي يمكن رؤيته عند مقارنة الشكل (19.12) مع الشكل (18.12)، كلتا السلسلتين تعطيان نفس الاتجاهات وهي حوالي 0.16 درجة مئوية/عِقْد لدرجات حرارة NCDC (أو NASA) خلال هذه الفترة أو 0.15 درجة مئوية/عِقْد لدرجات حرارة UAH.

الاستنتاجات من نماذج ARMAX البسيطة تتفق بشكل عام مع تلك التي تم الحصول عليها من تحليلات أكثر تعقيداً في دراسة Foster and Rahmstorf (2011)، حيث إن طريقتهما تتضمن البحث عن مواصفات فترة التباطؤ المثلى، وهذا على عكس ما قمنا به هنا باختيار عشوائي لفترة تباطؤ شهر واحد لكل المتغيرات التنبؤية. كما أن دراسة Foster and Rahmstorf تضمنت دوال مثلثية (سلاسل فوريير من الدرجة الثانية) لصياغة نماذج الدورات السنوية للبواقي في البيانات. كما تم تضمين الزمن بدلاً من تركيز CO_2 ضمن المتغيرات التنبؤية، وذلك لتوضيح الاتجاهات الزمنية. ولكن هذه الاتجاهات تم تفسيرها بأنها أسباب بشرية. ومبدئياً فإن هذه الأسباب البشرية يمكن أن تتضمن الغازات المسببة للاحتباس الحراري إلى جانب CO_2، والأنشطة الأخرى مثل إزالة الغابات، وتغير الغطاء النباتي، أو استنزاف طبقة الأوزون التي يمكن قياسها بواسطة متغير *co2anom*. ولذا فإن التفسير السببي لـ CO_2 قد يكون بسيطاً جداً. وعملياً فإن CO_2 ازداد بشكل مباشر مع الوقت (الشكل 10.12) مستخدماً الوقت أو CO_2 كمتغير تنبؤي سوف يعطي نتائج متشابهة.



الفصل الثالث عشر

صياغة نماذج التأثيرات المختلطة والمستويات المتعددة

Multilevel and Mixed-Effects Modeling

صياغة نماذج التأثيرات المختلطة هي عبارة عن تحليل انحدار يسمح بنوعين من التأثيرات، التأثيرات الثابتة: وهذا يعني أنها قيم التقاطع والميل، وهي تقوم بتفسير المجتمع ككل كما هو الوضع في الانحدار العادي. وأيضاً التأثيرات العشوائية: وتعني أن قيم التقاطع والميل يمكنها أن تكون متنوعة في المجموعات الفرعية للعينة. كل طرق الانحدار التي تم تناولها حتى الآن في هذا الكتاب، تتضمن التأثيرات الثابتة فقط. صياغة نماذج للتأثيرات المختلطة تفتح مجالاً جديدًا لاحتمالات نماذج المستويات المتعددة، وتحليل منحنى النمو، والبيانات الطولية، والسلاسل الزمنية المقطعية.

دراسة Albright and Marinova (2010) توضح مقارنة عملية بين طريقة صياغة النماذج المختلطة ببرنامج ستاتا وSAS وSPSS وR مع برنامج صياغة نموذج خطي ترتيبي the hierarchical linear modeling(HLM) تم تطويره بواسطة Raudenbush and Bryck (2002) and Raudenbush et al. (2005)، ويمكن الاطلاع على شرح أكثر تفصيلاً للنماذج الثابتة وارتباطاتها مع HLM في كتاب Rabe- Hesketh and Skrondal (2012)، وباختصار، فإن طرق HLM لصياغة النماذج ذات المستويات المتعددة تتم بعدة خطوات هي: تحديد معادلات منفصلة. فعلى سبيل المثال، لتأثيرات المستوى 1 والمستوى 2؛ وفي الواقع العملي هذه المعادلات لا يمكن تقديرها بطريقة منفصلة، ولذلك فإن البرنامج يقوم داخلياً بالاستعاضة عن ذلك من خلال إنشاء معادلة مختصرة واحدة للتقدير. طريقة صياغة نماذج التأثيرات المختلطة تعمل

مباشرة مع المعادلة المختصرة معطياً إياها مظهرًا ذا "مستويات متعددة أقل" من طريقة HLM حتى عندما تقوم الطريقتـان بتفسير النمـاذج المتكافئـة رياضياً. تأثيرات HLM عند مستويات مختلفة، يمكنها أن تكون ممثلة بشكل متكافئ للتأثيرات العشوائية أو الثابتة، مع معادلة مختصرة واحدة، وقد لاحظ Rabe-Hesketh and Skrondal (2012:171) بأن هناك بعض الاختلافات الثقافية بين HLM، وتطبيقات النماذج المختلفة: البحوث التي استخدمت [HLM] تميل لتضمين تفاعلات أكثر عند مستويات مختلفة، ومعاملات عشوائية أكثر فـي النماذج (لأن نماذج المستوى 2 تظهر بشكل غريب بدون بواقي) من البحوث التي تستخدم ستاتا مثلاً.

هناك ثلاثة أوامر ستاتا تمثل الأداة الأكثـر شـيوعاً لـصياغة نمـاذج التأثيرات المختلطة، وذات المستويات المتعـددة. الأمـر **xtmixed** يُناسـب النماذج الخطية مثل نظيره للتأثيرات المختلطة، وهو الأمر **regress**. وبالمثل، فإن الأمر **xtmelogit** يُناسب نماذج الانحدار اللوغاريتمية للتأثيرات المختلطة للمخرجات الثنائية، وكذلك فإن الأوامر **logit** و**logistic** و**xtmepoisson** تُناسب نماذج بواسون للتأثيرات المختلطة للمخرجات المعدودة مثل تعمـيم الأمـر **poisso**. برنامج ستاتا يوفر أيضاً عددًا أكثر من طـرق الحـساب المتعلقـة ببعض المهام المفاهيمية، الأمثلة تتضمن نماذج ذات حدين سالبة واحتمالية، ونماذج توبيت مع تقاطعات عشوائية. للحصول على قائمة كاملة مع روابط للتفاصيل عن كل أمر، قم بطباعة الأمر **help xt**. العديد مـن الأوامـر تـم تطويرها أولاً للاستخدام مع بيانات سلاسل زمنية مقطعية أو طولية، ولـذلك فإن أوامرها تبدأ بـ **xt**.

إجراءات الأوامر **xtmixed** و**xtmelogit** و**xtmepoisson** يمكن تطبيقها مـن خلال طباعة الأوامر، أو من خلال القوائم:

Statistics > Multilevel mixed-effects models

قوائم الحسابات الأخرى لـ **xt** تم وضعها بشكل منفصل تحت القائمـة التالية:

Statistics > Longitudinal/panel data

يحتوي دليل المستخدم *Longitudinal/Panel Data Reference Manual* على أمثلة، وتفاصيل تقنية، ومراجع للتأثيرات المختلطة، وطـرق **xt** الأخـرى، كتاب Luke (2004) يعرض مقدمة مختـصرة عـن صـياغة النمـاذج ذات المستويات المتعددة. كما أن هناك شرحاً أكثر تفصيلاً فـي كتـاب (2007) Bickel وكتاب McCulloch and Searle (2001) وكتاب Raudenbush and (2002) Bryk وكذلك كتاب Verbeke and Molenberghs (2000)، المرجع الخاص القيّم لمستخدمي برنامج ستاتا، هو كتاب بعنـوان *"Multilevel and Longitudinal Modeling Using Stata"* للمؤلفين Rabe-Hesketh and Skrondal (2012) يقـوم بشرح الطرق الرسمية لــ **xt** ببرنامج ستاتا وكذلك البرامج غيـر الرسـمية التي يُطلق عليها **gllamm** التي تقوم بإضافة قدرات لصياغة نمـاذج خطيـة عاملة للتأثيرات المختلطة، لمزيد من المعلومات عن كيفية الحصول وتثبيت الملفات التنفيذية ado-files لهذه البرامج قم بطباعة الأمر **findit gllamm**.

أمثلة عن الأوامر : Example Commands

```
.xtmixed crime year || city: year
```

يقوم بحساب انحدار التأثيرات المختلطة لمتغير *crime* (الجريمة) علـى متغير *year* (السنة) مع تقاطع عشوائي، وميل لكل قيمة من قيم المتغير *city* (المدينة)، ولذا سوف نحصل على معدلات الجريمة، والتي هي عبارة عـن تركيبة من الاتجاه العام (التأثيرات الثابتة) مع تباينات علـى ذلـك الاتجـاه (التأثيرات العشوائية) لكل مدينة.

```
.xtmixed SAT parentcoll prepcourse grades ||
 city:|| school: grades
```

يقوم بصياغة نموذج التأثيرات المختلطة متعدد المستويات أو الهرمـي متوقعاً نتائج SAT للطلبة كدالة لــ : (1) تأثيرات ثابتة أو تـأثيرات العينـة بالكامل لمعرفة ما إذا كان والد أو والديّ الطلبة متخرّج من كليـة، ومـا إذا كان الطالب قد أخذ دورة إعداد ومعدل الطالب، (2) تقاطعات عشوائية تمثل تأثير المدينة التي بها المدرسة، (3) تقاطع عشوائي وميل لمُعامِل معدل كل

طالب والذي قد يكون مختلفاً من مدرسة لأخرى؛ كل الطلبة (المشاهدات) لهم علاقة متشابكة بالمدارس، والأخيرة لها علاقة بالمدن التي تقع بهـا. لاحـظ الترتيب لأجزاء التأثيرات المختلطة في الأمر.

```
.xtmixed y x1 x2 x3 || state: x3
.estimates store A
.xtmixed y x1 x2 x3 || state:
.estimates store B
.lrtest A B
```

يقوم بحساب معدل الاحتمال لاختبار χ^2 لفرضية العدم التي تفترض بعدم وجود اختلاف في التناسب بين النموذج الأكثر تعقيداً *A*، والذي يتضمن ميلاً على المتغير *x3*، والنموذج الأبسط *B* (*B* موجود داخل *A*) وهو لا يتـضمن ميلاً عشوائياً على المتغير *x3*، هذا يصل لاختبار ما إذا كان الميل العشوائي على *x3* ذا معنوية إحصائية. الترتيب في النموذجين اللذين تم تحديدهما فـي الأمر **lrtest** ليس مهماً، إذا قمنا بطباعة **lrtest** ***B A*** فإن برنامج ستاتا سـوف يتعامل بشكل صحيح مع هذه العملية معتبراً أن *B* متداخل في *A*.

```
.xtmixed y x1 x2 x3 || state: x2 x3, reml
 nocons cov(unstructured)
```

يقوم بحساب انحدار التأثيرات المختلطة للمتغير *y* على التأثيرات الثابتة للمتغيرات التنبؤية *x1, x2, x3* وأيضاً على التأثيرات العشوائية للمتغيرات *x2,* *x3* لكل قيمة من قيم المتغير *state*، حيث يـتم الحـصول علـى التقـديرات بواسطة الحد الأقصى للاحتمال المحدود. النموذج يُفترض ألا يكون له تقاطع عشوائي، ولا مصفوفة تغاير غير منتظمة، والتي يكون بها تباينات التـأثير العشوائي، وقيم التغاير كلها مقدّرة بوضوح.

```
.estat recov
```

بعد الأمر **xtmixed** يقوم الأمر أعلاه بعرض مصفوفة التغاير- التبـاين المقدّرة للتأثيرات العشوائية.

```
.predict re*, reffects
```

بعد أمر التقدير **xtmixed** يقوم هذا الأمر بحساب أفضل تقديرات خطية غير المتحيزة (BLUPs) لكل التأثيرات العشوائية في النمـوذج. التـأثيرات العشوائية يتم حفظها كمتغيرات لها الأسماء التالية *re1, re2* وهكذا.

```
.predict yhat, fitted
```

بعد أمر التقدير **xtmixed** يقوم الأمر أعلاه بحساب القيم المتوقعة للمتغير *y*، وللحصول على التوقعات من جزء التأثيرات المختلطة للنموذج فقط نقوم بطباعة **predict yhat, xb**. الخيارات الأخرى **predict** تقوم بإيجاد الأخطـاء المعيارية للجزء الثابت (**stdp**) أو البواقي المعيارية (**rstan**). وللحصول على قائمة كاملة من أوامر ما بعد التقدير مع روابطها وكيفية تركيبها وخياراتها، قم بطباعة الأمر **help xtmixed postestimation**.

```
.xtmelogit y x1 x2 || state:
```

يقوم بحساب الانحدار اللوغاريتمي للتأثيرات الثابتة {0,1} للمتغير *y* علـى المتغيرين *x2, x1* مع التقاطعات العشوائية لكل مستوى من المتغير *state*.

```
.predict phat
```

بعد أمر التقدير **xtmelogit** يقوم الأمر أعلاه بحساب الاحتمالات المتوقعة من نموذج كامل (ثابت زائد عشوائي)، لمعرفة أوامر ما بعد التقدير الأخرى بالإضافة إلى قائمة كاملة بخيارات الأمر **predict** بما فيها بـواقي بواسـون (**pearson**) وبواقي الانحراف (**deviance**). للحصول على تفاصيل أكثر عـن أوامر ما بعد التقـدير وخياراتهـا، قـم بطباعـة الأمـر **help xtmelogit postestimation**.

```
.xtmepoisson accidents x1 x2 x3, exposure
 (persondays)|| season:|| port: , irr
```

يقوم بتقدير نموذج بواسون للتأثيرات المختلطة للمتغير *accidents* الذي يحسب حوادث قوارب صيد الأسماك. المتغيرات التنبؤية للتأثير الثابـت – وهي خصائص كل قارب صيد على حدة – هـي *x3, x2, x1*، ويـتم قيـاس التعرّض بعدد شخص/ يوم في البحر لذلك القارب، نقوم بتضمين التقاطعات العشوائية لكل موسم *season* أو سنة، والمدينة التي بها الميناء *port* التـي تتداخل مع المواسم، النتائج توضح مُعامِلات التأثير الثابت كنـسب لمعـدل الحوادث (**irr**).

```
.gllamm warming sex educ,i(region)
 family(binomial) link(ologit)adapt
```

يقوم بحساب النموذج الكامن الخطي المعياري والمختلط. وفي هذا المثال، يتم حساب الانحدار اللوغاريتمي المرتب للتأثيرات المختلطة للمتغير الترتيبي *warming* على المتغيرات التنبؤية للتأثير الثابت للمتغيرين *sex, educ*، ويتم إدراج التقاطعات العشوائية لكل قيمة من قيم المتغير *region* والتي تم تقديرها من خلال التربيع التكييفي. الخياران **family()** و**link()** يمكنهما تحديد النماذج الأخرى، بما فيها اللوغاريتم متعدد الحدود والاحتمال واللوغاريتم التكميلي. الأمر **gllamm** ليس من الأوامر الرسمية ببرنامج ستاتا، ولكنه متوافر على شبكة الإنترنت مجاناً، قم بطباعة الأمر **findit gllamm** للحصول على معلومات عن كيفية تحميله وتثبيته والملفات الضرورية الخاصة به، كتاب Rabe-Hesketh and Skrondal (2012) يعطي تفاصيل أكثر، وأمثلة عن كيفية استخدام الأمر **gllamm**.

الانحدار مع التقاطعات العشوائية :

Regression With Random Intercepts

لشرح الأمر **xtmixed** سوف نبدأ مع بيانات على مستوى الدولة، حيث إن البيانات تتعلق بالانتخابات الرئاسيّة بالولايات المتحدة لسنة 2004 (دراسة Robinson 2005). في هذه الانتخابات فاز جورج بوش (حصل على 50.7% من الأصوات) حيث هزم كلاً من جون كيري (48.3%) ورالف نادر (0.4%). السمة المثيرة في هذه الانتخابات كانت نمطها الجغرافي، حيث فاز كيري بولايات في الساحل الغربي والشمال الشرقي وحول البحيرات العظمى، بينما بوش فاز في كل الولايات الأخرى. دعم بوش كان قوياً في المناطق الريفية، بينما الأصوات التي حصل عليها كيري تركزت في المدن، ملف البيانات *election_2004i* يحتوي على نتائج الانتخابات والمتغيرات التي

تغطي مقاطعات الولايات المتحدة، هناك متغير نوعي يحتوي على تقسيمات السكان (*cendiv*) والذي يُقسّم الولايات المتحدة إلى 9 مناطق جغرافية. كما أن المتغيرات تتضمن العدد الكلي للأصوات (*votes*)، ونسبة أصوات بوش هـي (*bush*)، بينما لوغاريتم الكثافة السكانية (*logdens*) والذي يُعتبر كمؤشر على ريفية المنطقة، والمتغيرات الأخرى لنسبة سكان المقاطعة التي تنتمي إلى الأقليات العرقية (*minority*) أو البالغين الحاصلين على درجة جامعية (*colled*).

```
.use C:\data\election_2004i.dta، clear
.describe
```

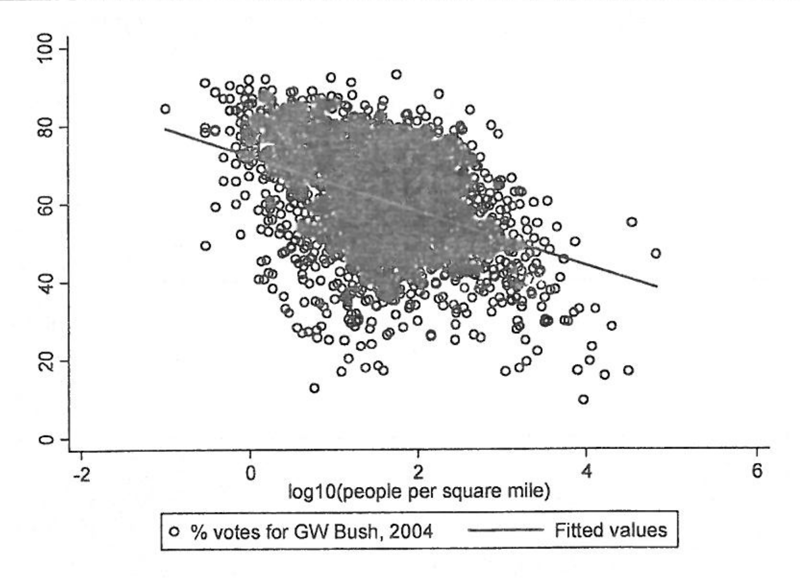
```
Contains data from C:\data\election_2004i.dta
  obs:         3,054                          US counties -- 2004 election (Robinson 2005)
 vars:            11                          2 Jul 2012 06:11
 size:       219,888
```

variable name	storage type	display format	value label	variable label
fips	long	%9.0g		FIPS code
state	str20	%20s		State name
state2	str2	%9s		State 2-letter abbreviation
region	byte	%9.0g	region	Region (4)
cendiv	byte	%15.0g	division	Census division (9)
county	str24	%24s		County name
votes	float	%9.0g		Total # of votes cast, 2004
bush	float	%9.0g		% votes for GW Bush, 2004
logdens	float	%9.0g		log10(people per square mile)
minority	float	%9.0g		% population minority
colled	float	%9.0g		% adults >25 w/4+ years college

Sorted by: fips

نسبة المقترعين لصالح بوش انخفضت كلما زادت كثافة السكان كما هو معروض في شكل الانتشار وخط الانحدار بالشكل (1.13)، كل نقطة تمثل مقاطعة واحدة من 3,054 مقاطعة.

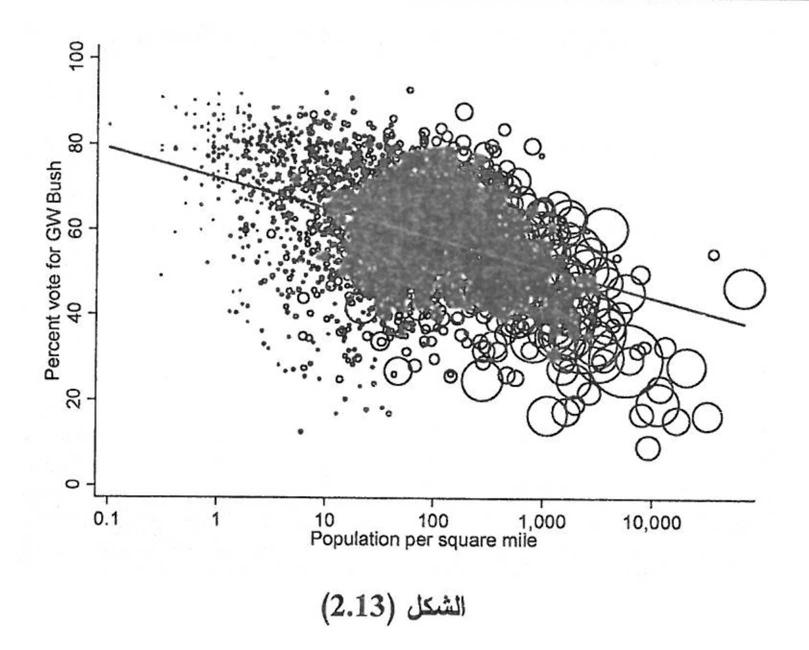
```
.graph twoway scatter bush logdens, msymbol(Oh)
    || lfit bush logdens, lwidth(medthick)
```



الشكل (1.13)

النسخة المطوّرة لهذا الشكل البياني الذي يمثل الكثافة السكانية مع الانتخابات تظهر في الشكل (2.13). القيم اللوغاريتمية على المحور الأفقي تم توصيفها (1 تصبح "10"، 2 تصبح "100" وهكذا) وذلك لجعل هذه القيم قابلة للقراءة بسهولة. وباستخدام *votes* كأوزان تكرارية لشكل الانتشار مما جعل العلامات بمنطقة شكل الانتشار جزءًا فقط إلى العدد الكلي للأصوات الانتخابية، حيث هناك فرق واضح بين المقاطعات التي تكون بها كثافة سكان كبيرة أو صغيرة، وإذا لم يتم ذلك فإن التحليل في هذا الفصل لن يستخدم الأوزان، سوف نركز هنا على اتجاهات التصويت بدلاً من الأفراد في هذه المقاطعات.

```
.graph twoway scatter bushlogdens [fw=votes],
 msymbol(Oh)
   || lfit bush logdens, lwidth(medthick)
   || , xlabel(-1 "0.1" 0 "1" 1 "10" 2 "100" 3
        "1,000" 4 "10,000", grid) legend(off)
   xtitle("Population per square mile")
   ytitle("Percent vote for GW Bush")
```



الشكل (2.13)

كما يؤكد الشكل (2.13) فنسبة التصويت لصالح جورج بوش تميل لتكون أقل في المناطق ذات الكثافة السكانية العالية والمقاطعات المتحضرة، كما أنها تميل أيضاً لتكون أقل في المقاطعات التي بها نسبة كبيرة من الأقليات العرقية أو نسبة أكبر من البالغين الذين يحملون مؤهلاً جامعياً.

.regress *bush logdens minority colled*

Source	SS	df	MS
Model	122345.617	3	40781.8725
Residual	358593.826	3037	118.075017
Total	480939.443	3040	158.203764

Number of obs	=	3041
F(3, 3037)	=	345.39
Prob > F	=	0.0000
R-squared	=	0.2544
Adj R-squared	=	0.2537
Root MSE	=	10.866

bush	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
logdens	-5.457462	.3031091	-18.00	0.000	-6.051781	-4.863142
minority	-.251151	.0125261	-20.05	0.000	-.2757115	-.2265905
colled	-.1811345	.0334151	-5.42	0.000	-.246653	-.115616
_cons	75.78636	.5739508	132.04	0.000	74.66099	76.91173

في صيغ النموذج المختلط، نقوم بتقدير نموذج يحتوي علـــى تـــأثيرات ثابتة فقط، ويكون التقاطع والمُعاملات تفسر العينة ككـــل. ونفـــس نمـــوذج التأثيرات الثابتة يمكن تقديرها باستخدام الأمر **xtmixed** مع نفـــس تركيبـــة الأمر السابق.

```
.xtmixed bush logdens minority colled

Mixed-effects ML regression                 Number of obs      =      3041

                                            Wald chi2(3)       =   1037.53
Log likelihood = -11567.783                 Prob > chi2        =    0.0000
```

bush	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
logdens	-5.457462	.3029097	-18.02	0.000	-6.051154	-4.86377
minority	-.251151	.0125179	-20.06	0.000	-.2756856	-.2266164
colled	-.1811345	.0333931	-5.42	0.000	-.2465838	-.1156852
_cons	75.78636	.5735732	132.13	0.000	74.66217	76.91054

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
sd(Residual)	10.85908	.1392419	10.58958	11.13545

الاحتمالية القصوى (ML) هي طريقة التقدير الافتراضية للأمر **xtmixed**، ولكن قد يتم تحديده بشكل خاص مع خيار **ml**، أو بشكل آخر فـــإن الخيـــار **reml** يمكن استخدامه لتقدير الاحتمالية القصوى المقيّدة، وللحصول على قائمة بالتقديرات والخيارات المتعلقة بها قم بطباعة الأمر **help xtmixed**.

النمط أو الاتجاه الجغرافي للتصويت يمكن رؤيته في الخرائط بـــألوان زرقاء وحمراء لهذه الانتخابات، والتي لم يـــتم حـــسابها بواســـطة نمـــوذج التأثيرات الثابتة أعلاه، والذي يفترض بأن التقاطع وقيم الميل هي نفسها لكل 3,041 مقاطعة في هذا التحليل، الطريقة الأخرى لصياغة نمـــوذج الاتجـــاه نحو نمط انتخاب مختلف في أجزاء مختلفة من المقاطعة (ولتخفيض مـــشكلة أخطاء الارتباط المكانية) تتم من خلال السماح لكـــل التقـــسيمات الـــسكانية التسعة لتأخذ تقاطعها العشوائي الخاص بها، وبدلاً مـــن نمـــوذج الارتبـــاط (التأثيرات الثابتة) المعتاد مثل:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \in_i \qquad [1.13]$$

يمكننا تضمين ليس فقط مجموعة مُعامِلات β التي تفسّر كل المقاطعات، وإنما أيضاً تقاطع عشوائي u_0 يتغير من تقسيم سكاني إلى آخر.

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \in_{ij} \qquad [2.13]$$

المعادلة [2.13] تصوّر قيمة المتغير y لكل مقاطعة i وكل تقسيم سكاني j كدالة لتأثيرات المتغيرات $x3, x2, x1$ والتي هي نفسها لكل التقسيمات. التقاطع العشوائي u_{0j} يسمح لاحتمالية أن تكون نسبة متوسط الانتخاب لجورج بوش أعلى أو أقل بانتظام بين المقاطعات لبعض التقسيمات، هـذا يبـدو مناسـباً لانتخابات الولايات المتحدة عند الأخذ بالاعتبار الأنماط الجغرافية الواضحة، كما يمكننا تقدير نموذج مع تقاطعات عشوائية لكل تقـسيم سـكاني، وذلـك بإضافة جزء من تأثيرات عشوائية جديدة إلى الأمر **xtmixed**:

```
.xtmixed bush logdens minority colled || cendiv:
Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log likelihood =  -11339.79
Iteration 1:   log likelihood =  -11339.79  (backed up)

Computing standard errors:

Mixed-effects ML regression                     Number of obs      =      3041
Group variable: cendiv                          Number of groups   =         9

                                                Obs per group: min =        67
                                                               avg =     337.9
                                                               max =       616

                                                Wald chi2(3)       =   1161.96
Log likelihood =  -11339.79                     Prob > chi2        =    0.0000

------------------------------------------------------------------------------
        bush |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    logdens |   -4.52417   .3621775   -12.49   0.000    -5.234025   -3.814316
   minority |  -.3645394   .0129918   -28.06   0.000    -.3900029   -.3390758
     colled |  -.0583942   .0357717    -1.63   0.103    -.1285053     .011717
      _cons |   72.09305    2.29404    31.43   0.000     67.59682    76.58929
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
cendiv: Identity             |
                   sd(_cons) |   6.617137   1.600468      4.119007    10.63036
-----------------------------+------------------------------------------------
               sd(Residual)  |   10.00339   .1284657      9.754742    10.25837
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =   455.99 Prob >= chibar2 = 0.0000
```

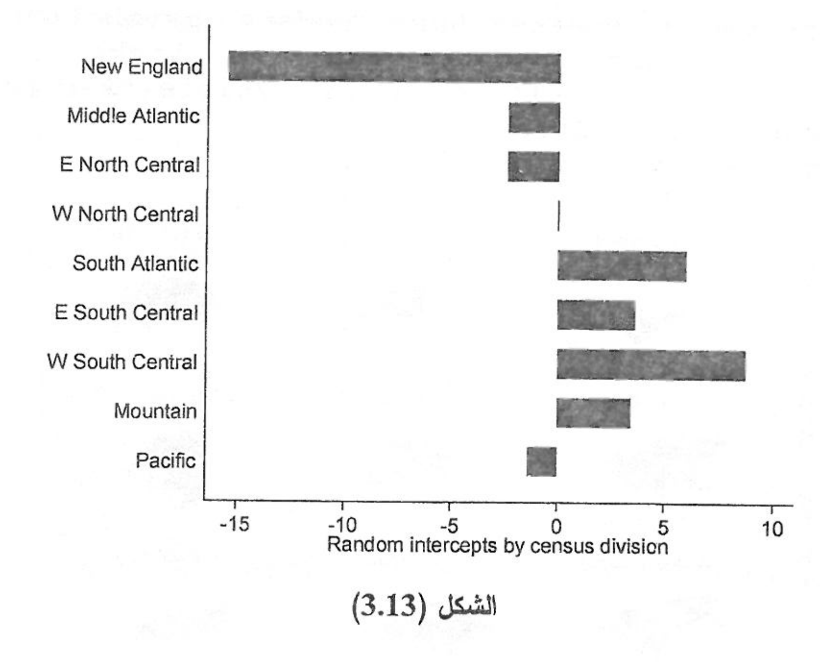
الجزء السابق من جدول مخرجات **xtmixed** يعرض جــزء التــأثيرات الثابتة للنموذج. هذا النموذج يقوم بتطبيق تسعة تقاطعات متفرقة، حيث يوجد هناك تقاطع لكل تقسيم سكاني، ولكن لم يتم تقديرها مباشرة، وبدلاً من ذلــك فإن الجزء السفلي من الجدول يعطي الانحراف المعياري المقدّر للتقاطعــات العشوائية (6.62) مع خطأ معياري (1.60) وفترة ثقة 95% لذلك الانحراف المعياري، وبذلك يصبح النموذج كما يلي:

$$bush_{ij} = 72.09 - 4.52logdens_{ij} - 0.36minority_{it} - 0.06colled_{ij} + u_{0j} + \in_{ij} \qquad [3.13]$$

إذا ظهر الانحراف المعياري لــ u_0 يختلف جوهرياً عن الصفر، فيمكننا أن نستنتج بأن هذه التقاطعات تتنوع من مكان لآخر، وهذا ما يبــدو عليــه الوضع هنا. حيث إن الانحراف المعياري أكبر من أربعة أخطاء معيارية من الصفر، وقيم هذا الانحراف جوهرية (6.62 نقطة مئوية) للمتغير التابع وهو نسبة التصويت لجورج بوش، اختبار نسبة الأرجحية العظمى يظهر في آخر سطر بجــدول المخرجات، والذي يؤكد بأن نمــوذج التقاطـــع العــشوائي شهــد تطــوّرًا كبيرًا عــن أي نموذج انحدار خطي مع تأثيرات ثابتة فقــط $(p \approx 0.0000)$.

بالرغم من أن الأمر **xtmixed** لا يحسب التأثيرات العــشوائية بطريقــة مباشرة، ولكن يمكننا الحصول على أفضل توقعات خطيــة غيــر متحيــزة (BLUPS) للتأثيرات العشوائية من خلال الأمر **predict**. الأوامر أدنــاه تقــوم بإنشاء متغير جديد باسم *randint0* يحتوي على التقاطعات العشوائية المتوقعة، ثم يقوم بإنشاء رسم بياني لكل تقاطع عشوائي مع كل تقسيم سكاني في الرسم البياني للأعمدة (الشكل 3.13).

```
.predict randint0, reffects
.graph hbar (mean) randint0, over(cendiv)
ytitle("Random intercepts by census division")
```



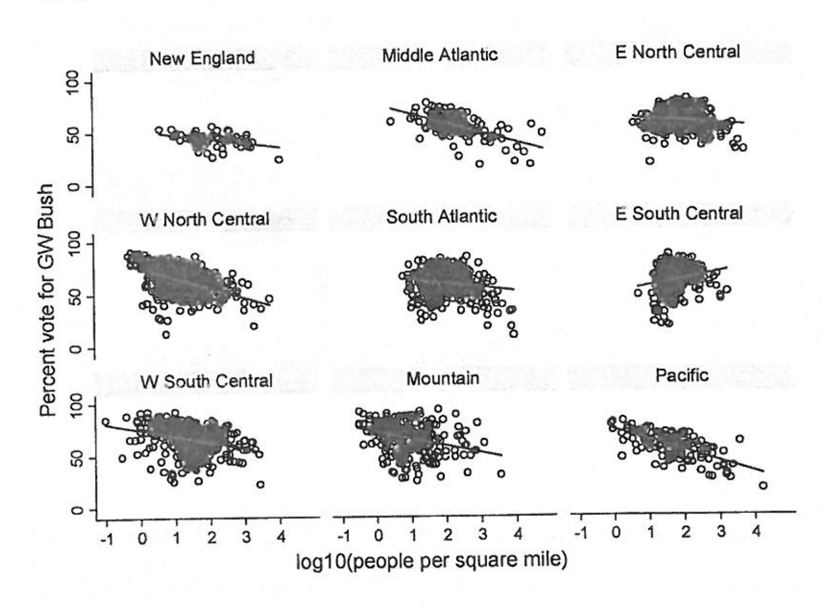
الشكل (3.13)

الشكل (3.13) يوضح بأنه عند أي مستوى من مستويات *logdens,* *minority, colled* نسبة الأصوات التي ذهبت لصالح بوش كانت بمتوسط 15 نقطة أقل في مقاطعات نيو إنجلاند New England وأكثر من 8 نقاط في المقاطعات الوسطى في الجنوب الغربي (مقاطعات Arkansas, Louisiana, Oklahoma, Texas) مقارنة مع منتصف الطريق وهو تقسيم المنطقة الوسطى للشمال الغربي.

التقاطعات والميول العشوائية : Random Intercepts and Slopes

في الشكل (2.13) رأينا – بأنه وبصفة عامة – أن نسبة الناخبين لصالح بوش تميل للانخفاض كلما زادت الكثافة السكانية. نموذج التقاطع العشوائي في الجزء السابق وافق على هذه النتيجة، وسمح للتقاطعات بأن تتغير حسب المناطق. ولكن ماذا لو كان ميل العلاقة بين الكثافة والانتخاب يتغير أيضاً بتغير المناطق؟ نظرة سريعة على شكل الانتشار لكل منطقة (الشكل 4.13) يعطينا مبررًا للشك بأن ذلك قد يحدث.

```
.graph twoway scatter bush logdens, msymbol(Oh)
    || lfit bush logdens, lwidth(medthick)
    ||  , xlabel(-1(1)4, grid) ytitle("Percent
    vote for GW Bush")
    by(cendiv, legend(off) note(""))
```



الشكل (4.13)

أصوات بوش تنخفض بدرجة كبيرة مع زيادة الكثافة السكانية في المناطق الوسطى والشمالية الغربية ومنطقة المحيط الهادئ. ولكن يبدو أن هناك علاقة ضعيفة في المناطق الوسطى الشمالية الشرقية، وحتى هناك تأثير موجب في المناطق الوسطى الجنوبية الشرقية. معامل التأثير الثابت السالب للمتغير *logdens* في النموذج السابق كان متوسطاً لهذه الاتجاهات المنخفضة والمستقرة والمرتفعة معاً.

يمكن إنشاء نموذج مختلط يتضمن قيم ميل عشوائية (u_{1j}) للمتغير التنبؤي x_1 والتقاطعات العشوائية (u_{0j}) لكل مجموعة من مجموعات j وتكون صيغته العامة كما يلي:

$$y_{ij}= \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + u_{0j} + u_{1j} x_{1j} + \epsilon_{ij} \quad [4.13]$$

ولتوضيح مثل هذا النموذج، سوف نقوم بإضافة متغير تنبؤي *logdens* لجزء التأثير المختلط بالأمر **xtmixed**.

```
.xtmixed bush logdens minority colled || cendiv:
 logdens
Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log likelihood = -11298.734
Iteration 1:   log likelihood = -11298.734

Computing standard errors:

Mixed-effects ML regression                     Number of obs      =      3041
Group variable: cendiv                          Number of groups   =         9

                                                Obs per group: min =        67
                                                               avg =     337.9
                                                               max =       616


                                                Wald chi2(3)       =    806.25
Log likelihood = -11298.734                     Prob > chi2        =    0.0000
```

bush	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
logdens	-3.310313	1.114965	-2.97	0.003	-5.495605	-1.125021
minority	-.3616886	.0130709	-27.67	0.000	-.387307	-.3360702
colled	-.1173469	.0360906	-3.25	0.001	-.1880833	-.0466105
_cons	70.12095	2.955209	23.73	0.000	64.32885	75.91305

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
cendiv: Independent				
sd(logdens)	3.113575	.8143897	1.86474	5.198768
sd(_cons)	8.5913	2.232214	5.162945	14.29619
sd(Residual)	9.825565	.1264176	9.580889	10.07649

```
LR test vs. linear regression:       chi2(2) =    538.10   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

وكما هو معتاد، فإن التأثيرات العشوائية لم يتم تقديرها مباشرة، وبدلاً من ذلك، فإن جدول مخرجات الأمر **xtmixed** يعرض تقديرات لانحرافاتها

المعيارية. الانحراف المعياري لمُعاملات لوغاريتم الكثافة هو 3.11 – وهو تقريباً أبعد أربع مرات من الأخطاء المعيارية (0.81) عن الصفر – مما يشير إلى وجود تباين ذي معنوية بين تقسيمات السكان في مُعاملات الميل، هذا الاستدلال سوف يتم دعمه بشكل حاسم من خلال اختبار معدل الأرجحية، ولإجراء هذا الاختبار، سوف نقوم بإعادة تقدير تقاطع النموذج، ويتم حفظ هذا التقدير باسم *A* (اسم تم اختياره عشوائياً) ثم نقوم بإعادة تقدير ميل وتقاطع النموذج وحفظ هذه التقديرات باسم *B*، وأخيراً نقوم بحساب اختبار معدل الأرجحية العظمى لمعرفة ما إذا كان *B* يتناسب بدرجة أفضل من تناسب *A*، وهذا ما حدث فعلاً في هذا المثال ($p \approx 0.0000$). ولذا فإننا نستنتج بأن إضافة قيم ميل عشوائية يقوم بتطوير النموذج بصورة كبيرة.

```
.quietly xtmixed bush logdens minority colled
  || cendiv:
.estimates store A
.quietly xtmixed bush logdens minority colled
  || cendiv: logdens
.estimates store B
.lrtest A B

Likelihood-ratio test                         LR chi2(1)  =     82.11
(Assumption: A nested in B)                   Prob > chi2 =    0.0000

Note: The reported degrees of freedom assumes the null hypothesis is not on the boundary of the
      parameter space.  If this is not true, then the reported test is conservative.
```

مخرجات الأمر **lrtest** تحذرنا بأن هناك "افتراضاً أن كل فرضيات العدم ليست ضمن حدود مساحة المعلمية". وبخلاف ذلك، فإن الاختبار الناتج سوف يكون محافظاً على هذه الحدود. كلتا الملاحظتين تشيران إلى نفس القضية الإحصائية، فالتباين لا يمكن أن يكون أقل من الصفر. ولذا فإن فرضية العدم القائلة بأن التباين يساوي صفراً تقع ضمن حدود المساحة المعلمية. في هذه الحالة الاحتمالية التي نتجت من اختبار معدل الأرجحية العظمى تمثل حداً أعلى "محافظاً على حدود" الاحتمال الواقعي، الأمر **xtmixed** يكتشف هذا الوضع بشكل تلقائي، وهذا لا يمكن القيام به باستخدام الأمر **lrtest**.

النموذج السابق يفترض أن قيم الميل والتقاطعات العشوائية غير مترابطة، وهذا يُكافئ إضافة الخيار cov(independent) الذي يُحدد تركيبة التباين المصاحب. الاحتمالات الأخرى تتضمن cov(unstructured) والذي يسمح بتوضيح التباين المصاحب الذي لايساوي صفرًا بين التأثيرات العشوائية.

```
.xtmixed bush logdens minority colled
  || cendiv: logdens, cov(unstructured)

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log likelihood =  -11296.31
Iteration 1:   log likelihood =  -11296.31  (backed up)

Computing standard errors:

Mixed-effects ML regression                     Number of obs      =      3041
Group variable: cendiv                          Number of groups   =         9

                                                Obs per group: min =        67
                                                               avg =     337.9
                                                               max =       616


                                                Wald chi2(3)       =    799.68
Log likelihood =  -11296.31                     Prob > chi2        =    0.0000
```

bush	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
logdens	-3.150009	1.169325	-2.69	0.007	-5.441844	-.858175
minority	-.3611161	.0130977	-27.57	0.000	-.3867872	-.3354451
colled	-.1230445	.0361363	-3.41	0.001	-.1938704	-.0522186
_cons	69.85194	3.168479	22.05	0.000	63.64184	76.06204

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
cendiv: Unstructured				
sd(logdens)	3.282749	.8547255	1.970658	5.468447
sd(_cons)	9.240389	2.402183	5.551459	15.3806
corr(logdens,_cons)	-.675152	.1958923	-.909687	-.1140965
sd(Residual)	9.823658	.1263468	9.579118	10.07444

```
LR test vs. linear regression:      chi2(3) =    542.95   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

الترابط المقدّر بين الميل العشوائي للمتغير *logdens* والتقاطع العـشوائي هو 0.675− وهو يبعد عن الصفر بأكثر من ثلاثة أخطاء معيارية. اختبـار معدل الأرجحية العظمى يتفق مع القول بأن السماح بهذا الترابط ينتج عنـه تطوّر ذو معنوية ($p = 0.277$) في النموذج الحالي عن النموذج السابق.

```
.estimates store C
.lrtest B C
 Likelihood-ratio test                              LR chi2(1)  =      4.85
 (Assumption: B nested in C)                        Prob > chi2 =    0.0277
```

النموذج الحالي هو:

$$bush_{ij} = 69.85 - 3.15\ logdens_{ij} - 0.36\ minority_{ij} - 0.12\ colled_{ij} + u_{0j} + u_{1j}logdens_{ij} + \in_{ij} \quad [5.13]$$

إذن ما الذي تقوم به قيم الميل لربط التصويت مع الكثافة لكـل تقـسيم سكاني؟ مرة أخرى يمكننا الحصول على قيم التأثيرات العشوائية (تم تسميتها هنا *randint1* و *randslo1*) من خلال الأمر **predict**، فالبيانـات التـي نقـوم باستخدامها الآن أصبحت تحتوي على مجموعة متغيرات جديدة.

```
.predict randslo1randint1, reffects
.describe
Contains data from C:\data\election_2004i.dta
  obs:         3,054                          US counties -- 2004 election (Robinson 2005)
 vars:            17                          23 Feb 2014 01:37
 size:       265,698
-------------------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------------------
fips            long   %9.0g                  FIPS code
state           str20  %20s                   State name
state2          str2   %9s                    State 2-letter abbreviation
region          byte   %9.0g       region     Region (4)
cendiv          byte   %15.0g      division   Census division (9)
county          str24  %24s                   County name
votes           float  %9.0g                  Total # of votes cast, 2004
bush            float  %9.0g                  % votes for GW Bush, 2004
logdens         float  %9.0g                  log10(people per square mile)
minority        float  %9.0g                  % population minority
colled          float  %9.0g                  % adults >25 w/4+ years college
randint0        float  %9.0g                  BLUP r.e. for cendiv: _cons
_est_C          byte   %8.0g                  esample() from estimates store
_est_A          byte   %8.0g                  esample() from estimates store
_est_B          byte   %8.0g                  esample() from estimates store
randslo1        float  %9.0g                  BLUP r.e. for cendiv: logdens
randint1        float  %9.0g                  BLUP r.e. for cendiv: _cons
-------------------------------------------------------------------------------------------
Sorted by:  fips
     Note:  dataset has changed since last saved
```

مُعاملات الميل العشوائي تتراوح من 5.45− للمقاطعات في تقسيم المنطقة الوسطى الشمالية الغربية W North Central إلى 4.49+ في المنطقة الوسطى الجنوبية الشرقية E South Central.

```
.table cendiv, contents(mean randslo1 mean
 randint1)
```

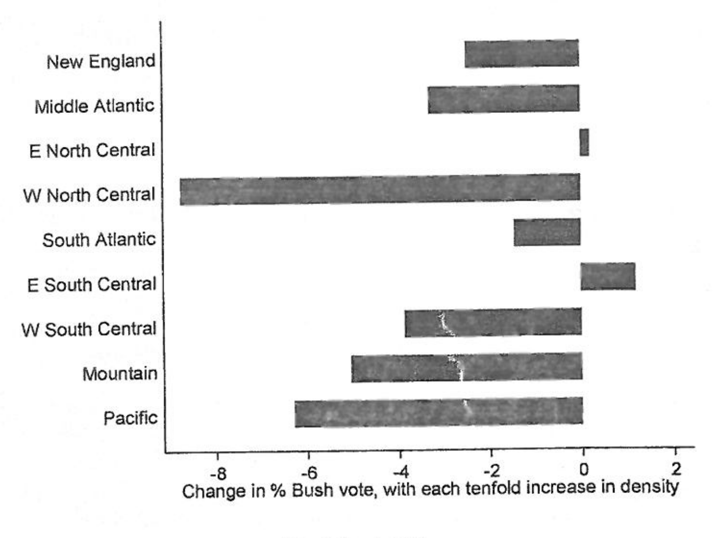
Census division (9)	mean(randslo1)	mean(randint1)
New England	.8396833	-16.95975
Middle Atlantic	.0018003	-2.339407
E North Central	3.510326	-8.810173
W North Central	-5.453928	7.912671
South Atlantic	1.870723	2.622947
E South Central	4.493841	-4.029463
W South Central	-.5490148	10.37175
Mountain	-1.73396	6.80135
Pacific	-2.979471	4.430075

لتوضيح العلاقة بين التصويت والكثافة السكانية يمكننا إعادة تنظيم المعادلة [5.13] بدمج قيم الميل العشوائي والثابت للمتغير *logdens*:

$$bush_{ij} = 69.85 + (u_{1j} - 3.15)\, logdens_{ij} - 0.36 minority_{ij} - 0.12 colled_{ij} + u_{0j} + \epsilon_{ij} \quad [6.13]$$

بعبارة أخرى، الميل لكل تقسيم سكاني يساوي ميل التأثير الثابت للعينة بالكامل زائداً ميل التأثير العشوائي لكل تقسيم سكاني. فمثلاً ضمن المقاطعات المطلّة على المحيط الهادئ نجد أن قيمة الميل الموحد هي 6.37- = 3.22 − 3.15-، قيم الميل الموحّدة التسعة تم حسابها وتمثيلها بيانياً في الشكل (5.13).

```
.gen slope1 = randslo1 + _b[logdens]
.graph hbar (mean) slope1, over(cendiv)
 ytitle("Change in % Bush vote, with each
 tenfold increase in density")
```



الشكل (5.13)

الشكل (5.13) يعرض كيف أن التدرج بين الريف والحضر في السلوك الانتخابي يختلف من مكان لآخر، في المقاطعات الوسطى الشمالية الغربية والمناطق الجبلية والمطلة على المحيط الهادئ كانت نسبة المصوّتين لبوش انخفضت بشكل حاد كلما زادت الكثافة السكانية، أما في المناطق الوسطى الشمالية الشرقية والوسطى الجنوبية الشرقية فقد كانت النسبة في الاتجاه الآخر، حيث زادت أصوات بوش بدرجة بسيطة، قيم الميل الموحدة في الشكل (5.13) تشابه بشكل عام تلك التي تظهر في الشكل (4.13) ولكن لا تتطابق بالضبط لأن قيم الميل الموحدة (المعادلة [6.13] أو الشكل 5.13) تم أيضاً تعديلها بالنسبة لتأثير الأقليات وخريجي الجامعات. في الجزء التالي سوف نقوم بدراسة ما إذا كانت هذه التأثيرات لها مكونات عشوائية.

قيم الميل العشوائية المتعددة : Multiple Random Slopes

لتحديد المُعاملات العشوائية للمتغيرات *logdens, minority, colled* يمكننا ببساطة إضافة أسماء هذه المتغيرات إلى جزء التأثيرات العشوائية للأمر **xtmixed**، لغرض إجراء اختبارات المقارنة. لاحقاً سوف نقوم بحفظ نتائج

التقدير باسم *full*، وهناك بعض التفاصيل المكررة تم إهمالها في المخرجات التالية:

```
.xtmixed bushlogdens minority colled
 || cendiv: logdens minority colled
Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log likelihood = -11184.804
Iteration 1:   log likelihood = -11184.804

Computing standard errors:

Mixed-effects ML regression                   Number of obs      =      3041
Group variable: cendiv                        Number of groups   =         9

                                              Obs per group: min =        67
                                                             avg =     337.9
                                                             max =       616


                                              Wald chi2(3)       =     52.49
Log likelihood = -11184.804                   Prob > chi2        =    0.0000
```

bush	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
logdens	-2.717128	1.373684	-1.98	0.048	-5.409499	-.0247572
minority	-.3795605	.0560052	-6.78	0.000	-.4893286	-.2697924
colled	-.1707863	.1727742	-0.99	0.323	-.5094175	.167845
_cons	70.86653	3.435918	20.63	0.000	64.13225	77.6008

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
cendiv: Independent				
sd(logdens)	3.868421	.9832899	2.350564	6.366421
sd(minority)	.153172	.0439569	.0872777	.2688161
sd(colled)	.5032414	.1241234	.310334	.8160625
sd(_cons)	10.01157	2.547813	6.079707	16.48625
sd(Residual)	9.375994	.1209753	9.141859	9.616124

```
LR test vs. linear regression:       chi2(4) =    765.96   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

```
.estimates store full
```

باعتبار نموذج *full* كخط أساسي، فــإن اختبــارات معــدل الأرجحيــة العظمى تُحدد بأن المُعاملات العشوائية للمتغيرات *logdens, minority, colled* كل منها لها اختلاف ذو معنوية إحصائية. ولذا فإن هذه الاختلافــات يجــب الاحتفاظ بها بالنموذج، فمثلاً لتقييم التأثيرات العشوائية للمتغير *colled*، فإننا نقوم بتقدير نموذج جديد بدونها (*nocolled*) ثم نقارن ذلك النموذج مع نموذج *full*، أما نموذج *nocolled* يتناسب بطريقة أسوأ بكثير ($p \approx 0.0000$) من نموذج *full* الذي رأيناه سابقاً.

```
.quietly xtmixed bush logdens minority colled
 || cendiv: logdens minority
.estimates store nocolled
.lrtest nocolled full

Likelihood-ratio test                                 LR chi2(1)  =    197.33
(Assumption: nocolled nested in full)                 Prob > chi2 =    0.0000

Note: The reported degrees of freedom assumes the null hypothesis is not on the boundary of the
      parameter space.  If this is not true, then the reported test is conservative.
```

خطوات متشابهة مــع نمــوذجين آخــرين (*nologdens* و *nominority*). واختبارات معدل الأرجحية العظمى توضح بأن نموذج *full* أيــضاً يتناســب بشكل أفضل بكثير من النماذج التي كانت بدون مُعامِل عــشوائي للمتغيــر *logdens* أو بدون مُعامِل عشوائي للمتغير *minority*.

```
.quietly xtmixed bush logdens minority colled
 || cendiv: minority colled
.estimates store nologdens
.lrtest nologdens full

Likelihood-ratio test                                 LR chi2(1)  =    124.87
(Assumption: nologdens nested in full)                Prob > chi2 =    0.0000

Note: The reported degrees of freedom assumes the null hypothesis is not on the boundary of the
      parameter space.  If this is not true, then the reported test is conservative.
```

يمكننا التحقق من أن تفاصيل كل هذه التأثيرات العشوائية أو التــأثيرات المدمجة التي ظهرت في النتائج من خلال حسابات وخطوط لتلك التــي تــم عرضها سابقاً في الشكل (5.13).

بحوث النماذج المختلطة في العادة تركز على التـــأثيرات الثابتـــة مـــع التأثيرات العشوائية والتي يتم إدراجها لتمثل عدم التجانس heterogeneity في البيانات وليس لأي أسباب موضوعية أخرى. فعلى سبيل المثـــال، التحاليـــل التي قمنا بها حتى الآن أظهرت بأن الكثافة السكانية ونسبة الأقليات ونـــسبة خريجي الجامعات يُمكنها أن تتنبأ بنمط التصويت في المقاطعة حتـــى بعـــد الأخذ في الاعتبار الاختلافات الإقليمية في متوسط الأصوات والاختلافـــات الإقليمية في الكثافة وخريجي الجامعات. ومن ناحية أخرى، فإن التـــأثيرات العشوائية هي نفسها لها أهمية. ولمعرفة كيف أن العلاقـــة بـــين التـــصويت ونسبة خريجي الجامعات (نسبة الأقليات أو لوغاريتم الكثافة السكانية) يختلف بين التقسيمات السكانية فيمكننا أن نتوقع التأثيرات العشوائية من خلال حساب التأثير الكلي. هذه الخطوات تم شرحها للتأثير الكلي لمتغيـــر *colled* علـــى النموذج أدناه بالكامل، وتمثيلها بيانياً في الشكل (6.13).

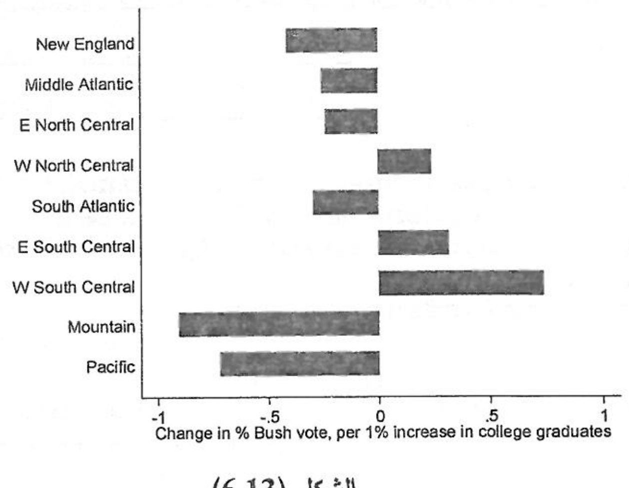
```
.quietly xtmixed bush logdens minority colled
 || cendiv: logdens minority colled
.predict relogdens reminority recolled re_cons,
 reffects
.describe relogdens-re_cons
```

variable name	storage type	display format	value label	variable label
relogdens	float	%9.0g		BLUP r.e. for cendiv: logdens
reminority	float	%9.0g		BLUP r.e. for cendiv: minority
recolled	float	%9.0g		BLUP r.e. for cendiv: colled
re_cons	float	%9.0g		BLUP r.e. for cendiv: _cons

```
.generate tecolled = recolled + _b[colled]
.label variable tecolled "random + fixed effect
 of colled"
.table cendiv, contents(mean recolled mean
 tecolled)
```

Census division (9)	mean(recolled)	mean(tecolled)
New England	-.2411761	-.4119623
Middle Atlantic	-.0849539	-.2557401
E North Central	-.0682224	-.2390087
W North Central	.4122158	.2414296
South Atlantic	-.1243222	-.2951084
E South Central	.4856058	.3148195
W South Central	.9054216	.7346353
Mountain	-.7355713	-.9063575
Pacific	-.5489974	-.7197837

```
.graph hbar (mean) tecolled, over(cendiv)
 ytitle("Change in % Bush vote, per 1%
 increase in college graduates")
```



الشكل (6.13)

الشكل (6.13) يعرض بيانياً سبب التطور الكبير الذي حدث في النموذج ($p \approx 0.0000$) بعد إضافة ميل عشوائي للمتغير *colled*، التأثيرات الكلية للمتغير *colled* على أصوات بوش تتراوح من سالب بدرجة كبيرة (نسبة أصوات بوش منخفضة في المقاطعات التي بها خريجو جامعات أكثر) في الجبـل Mountain (0.91−) والمحيط الأطلسي Pacific (0.72−) ذات الكثافة السكانية إلى بسيطة جـداً وغير ملحـوظة في مناطق جنوب المحيط الأطلسي South Atlantic أو المنطقة الوسطى الشمالية الغربية W North Central ثم إلى موجبة بدرجة كبيرة

(0.73+) في المناطق الوسطى الجنوبية الغربية W South Central حيث أصوات بوش كانت مرتفعة بدرجة كبيرة في تلك المقاطعات التي بها خريجو جامعات أكثر، مع ملاحظة التحكم في الكثافة السكانية ونسبة الأقليات والتأثيرات الإقليمية الأخرى، إذا قمنا بتقدير تأثير ثابت للمتغير *colled* فإن النموذج سوف يقوم بكفاءة بحساب متوسط التأثيرات السالبة التي تقترب من الصفر، والتأثيرات العشوائية الموجبة للمتغير *colled* في معامل ثابت موجب أسبوعي واحد، ويكون المعامل بالضبط 0.18- في انحدارات التأثيرات الثابتة الاثنين التي بدأنا بها في هذا الفصل.

الأمثلة التي رأيناها حتى الآن تم التعامل معها بشكل ناجح بواسطة الأمر **xtmixed**، ولكن هذا ليس هو الوضع دائماً، فتقدير نموذج ثابت يمكن أن يفشل في المعالجة لعدة أسباب مؤدياً إلى تكرار"عدم الانحناء" أو "التراجع" أو رسائل خطأ حول Hessian أو حسابات الأخطاء المعيارية. دليل المستخدم *Longitudinal/Panel Data Reference Manual* يُناقش كيفية تشخيص ومعالجة مشاكل التقارب convergence، والسبب المتكرر يبدو هو اقتراب مكونات تباين النماذج من الصفر مثل المعاملات العشوائية التي لا تتباين بدرجة كبيرة أو لها تباين منخفض. في مثل هذه الحالات، فإن المكونات المسببة لهذه المشكلة يمكن استبعادها بدرجة معقولة.

المستويات المتشابكة : Nested Levels

نماذج التأثيرات الثابتة يمكن تضمينها أكثر من مستوى متشابك واحد. فمثلاً المقاطعات في بيانات الانتخابات ليست متشابكة فقط مع تقسيمات السكان، ولكنها أيضاً مع الولايات والتي هي متشابكة مع تقسيمات السكان. هل التأثيرات العشوائية موجودة فقط عند مستوى تقسيمات السكان وأيضاً عند مستويات أقل من الولايات؟ الأمر **xtmixed** يسمح بمثل هذه النماذج الهرمية. فأجزاء التأثيرات العشوائية الإضافية تم إضافتها للأمر مع وحدات (متشابكة) أصغر على التوالي إلى اليمين. التحليل التالي يحدد التقاطعات العشوائية، وقيم الميل في المتغيرات التنبؤية الثلاثة لكل تقسيم سكاني، وأيضاً التقاطعات العشوائية، وقيم الميل لنسبة خريجي الجامعات *colled* لكل ولاية.

```
.xtmixed bush logdens minority colled
```

```
|| cendiv: logdens minority colled
|| state: colled
Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log likelihood = -10719.828
Iteration 1:   log likelihood = -10719.821
Iteration 2:   log likelihood = -10719.821

Computing standard errors:

Mixed-effects ML regression                     Number of obs      =      3041
```

Group Variable	No. of Groups	Observations per Group Minimum	Average	Maximum
cendiv	9	67	337.9	616
state	49	1	62.1	254

```
                                                Wald chi2(3)       =     68.85
Log likelihood = -10719.821                     Prob > chi2        =    0.0000
```

bush	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
logdens	-2.473536	.996434	-2.48	0.013	-4.426511	-.5205616
minority	-.4067648	.0533714	-7.62	0.000	-.5113709	-.3021588
colled	-.1787849	.1298313	-1.38	0.168	-.4332495	.0756797
_cons	71.13208	3 048286	23.34	0.000	65.15755	77.10661

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
cendiv: Independent				
sd(logdens)	2.703845	.7727275	1.544252	4.734186
sd(minority)	.1465435	.0428326	.0826365	.2598728
sd(colled)	.3683903	.0962733	.220729	.6148326
sd(_cons)	8.416873	2.417524	4.793643	14.77869
state: Independent				
sd(colled)	.1305727	.039009	.0727032	.2345047
sd(_cons)	5.883451	.7431715	4.593166	7.536196
sd(Residual)	7.863302	.1027691	7.664436	8.067328

```
LR test vs. linear regression:     chi2(6) =  1695.92   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

عند إلقاء نظرة سريعة على كل التأثيرات العشوائية لكلا المتغيرين، فإن المتغيرين *cendiv, state* في المخرجات أعلاه يبدو أنهما ذات معنوية، وهـذا يبدو واضحاً من خلال فترات الثقة والأخطاء المعيارية لهـذين المتغيـرين. الانحراف المعياري للمعاملات العشوائية عند مـستوى الولايـات للمتغيـر *colled* (0.13) أقل من الانحراف المعياري للمعاملات العشوائية عند مستوى تقسيمات السكان (0.37) ولكن كلاهما متعلق بدرجة كبيرة بمعامل التـأثير العشوائي للمتغير *colled* (0.18−)، أما فترة الثقة لمعامِل مـستوى الولايـة فيتراوح ما بين 0.07 إلى 0.23، واختبار معدل الأرجحية العظمى يُشير إلى أن هذا النموذج (والذي يُسمى هنا *state*) مع تقاطعات عـشوائية لمـستوى الولاية وقيم ميل أكثر تناسباً من النموذج الـسابق (*full*) والـذي كـان لـه تقاطعات عشوائية لمستوى تقسيمات السكان وقيم الميل.

```
.estimates store state
.lrtest full state
```

```
Likelihood-ratio test                                 LR chi2(2)  =    929.97
(Assumption: full nested in state)                    Prob > chi2 =    0.0000

Note: The reported degrees of freedom assumes the null hypothesis is not on the boundary of the
      parameter space.  If this is not true, then the reported test is conservative.
```

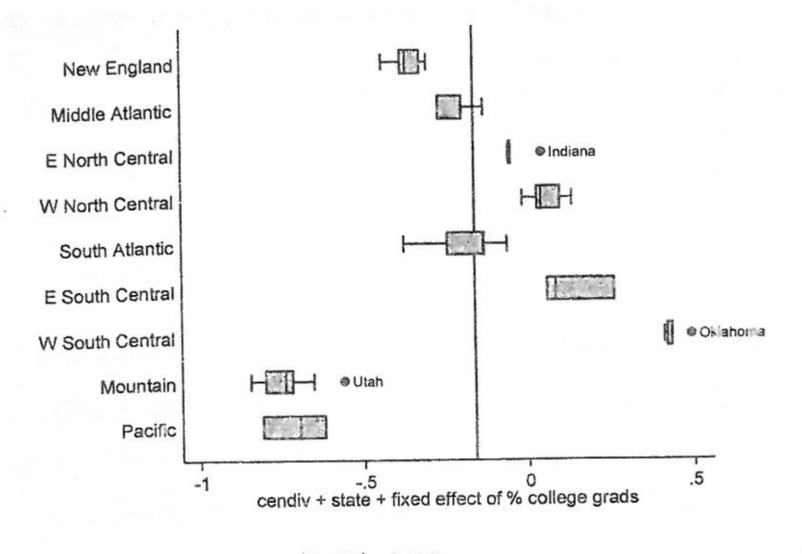
وكما تم سابقاً، فإنه يمكننا أن نتوقع التأثيرات العشوائية ثم نـستخدمها لحساب ورسم التأثيرات الكلية، بالنسبة للمتغير *colled* لـدينا الآن تـأثيرات عشوائية من 49 ولاية مختلفة، الرسم البياني الـصندوقي يمثـل التوزيـع بطريقة جيدة (الشكل 7.13) والذي يتبع نمطاً عاماً لتأثيرات تقسيم الـسكان الذي رأيناه سابقاً في الشكل (6.13) ولكنه الآن مع تباين لكل تقـسيم. ففـي الرسم البياني تظهر إنديانا (المنطقة الوسطى الشمالية الشرقية) وأوكلاهومـا (المنطقة الوسطى الجنوبية الغربية) كقيم متطرفة، لأن كلاً منهمـا يظهـر بصورة غير معتادة في تقسيمه الخاص.

```
.predict re*, reffects
.describe re1-re6
```

```
                  storage  display     value
variable name       type   format      label      variable label

re1               float   %9.0g                  BLUP r.e. for cendiv: logdens
re2               float   %9.0g                  BLUP r.e. for cendiv: minority
re3               float   %9.0g                  BLUP r.e. for cendiv: colled
re4               float   %9.0g                  BLUP r.e. for cendiv: _cons
re5               float   %9.0g                  BLUP r.e. for state: colled
re6               float   %9.0g                  BLUP r.e. for state: _cons
```

```
.gen tecolled2 = re3 + re5 + _b[colled]
.label variable tecolled2
 "cendiv + state + fixed effect of % college
 grads"
.graph hbox tecolled2, over(cendiv) yline(-.16)
 marker(1, mlabel(state))
```



الشكل (7.13)

ولمعرفة أدوات ما قبل التقدير الأخرى **xtmixed** قم بطباعة الأمر **help xtmixed postestimation**، كما أن دليل المستخدم *Longitudinal/Panel Data Reference Manual* وكتاب Rabe-Hesketh and Skrondal (2012) تعرض تطبيقات عن صياغة النماذج المختلطة مثل هياكل التغاير القُطريّة المقفلة، ونماذج التأثيرات المقطعية.

المقاييس المتكررة : Repeated Measurements

الملف *attract2.dta* يحتوي على بيانات لتجربة غريبة تم إجراؤها في احتفال لطلبة بالجامعة، وخلال هذا الحفل يقوم الطلبة بتناول بعض المشروبات، في هذه التجربة تم سؤال 15 طالباً جامعياً ليقوموا بتقييم فردي لجاذبية صور رجال ونساء لايعرفونهم على مقياس من 1 إلى 10، تم تكرار عملية التقييم لكل مشارك من خلال إعطائه نفس الصور بعد خلطها عشوائياً أربع مرات خلال فترة المساء. المتغير *ratemale* يمثل متوسط التقييم الذي أعطاه كل مشارك لكل صور الذكور في جلسة واحدة. والمتغير *ratefem* يمثل متوسط التقييم الذي تم إعطاؤه لصور الإناث. المتغير *gender* يُسجل جنس المشارك نفسه، والمتغير *bac* يسجل نسبة الكحول في الدم والذي يُقاس بجهاز قياس الكحول في الدم.

```
.use C:\data\attract2.dta, clear
.describe
```

```
Contains data from C:\data\attract2.dta
  obs:           204                          Perceived attractiveness and drinking -- DC
                                                Hamilton (2003)
 vars:             7                          2 Jul 2012 06:11
 size:         3,876
```

variable name	storage type	display format	value label	variable label
id	byte	%9.0g		Participant number
gender	byte	%9.0g	gender	Gender
bac	float	%9.0g		Blood alchohol content
single	byte	%9.0g	single	Relationship status single
drinkfrq	float	%9.0g		Days drinking in previous week
ratefem	float	%9.0g		Rated attractiveness of females
ratemale	float	%9.0g		Rated attractiveness of males

Sorted by: id bac

فرضيات البحث تتضمن أن "أكواب الجُعّة" لها تأثير على تقييم المشارك: هل الغرباء يُصبِحون أكثر جاذبية إذا قام المشارك بشرب كميات أكبر من الكحول؟ وفي سياق هذه التجربة، هل هناك علاقة إيجابية بين نسبة

الكحول في الدم، ومعدلات الجاذبية المعطاة للصور؟ وإذا كان الأمر كذلك، فهل هذه العلاقة تختلف بين جنسي المشاركين أو بين الصور؟

بالرغم من أن البيانات تحتوي على 204 مشاهدات، فإنها تمثل 51 مشاركاً فقط، ويبدو أنه من المعقول الاعتقاد بأن المشاركين ربما يختلفون في ميولهم لإعطاء معدلات أعلى أو أقل نوعاً ما، ويختلفون في ردة فعلهم تجاه الكحول. نموذج تأثيرات مختلطة مع تقاطعات عشوائية وقيم ميل يمكنه استيعاب هذه التعقيدات المحتملة. هذا الوضع يختلف عن مثال الانتخابات السابق في أن التقاطعات العشوائية الفردية وقيم الميل لن تكون مثيرة للاهتمام بدرجة كبيرة، لأنها تفسر أفرادًا مجهولين. أما هذه التجربة فهي عبارة عن بيانات أو ملامح تصميم تجريبي نحتاج لتعديله عند اختبار الفرضيات الرئيسة.

باستخدام رموز المتغير العاملي، فإن الأمر **xtmixed** أدناه يُحدد نموذجاً به متوسطات التقييم المعطاة لصور وجوه الإناث (*ratefem*) ويتم التنبؤ بها بواسطة التأثيرات الثابتة للمتغير الإشاري *gender*، والمتغير المستمر *bac* وتفاعلهما، بالإضافة إلى ذلك، فإن النموذج يتضمن تقاطعات عشوائية، وقيم ميل للمتغير *bac* والتي يمكن أن تختلف بين المشاركين.

```
.xtmixed ratefem i.gender##c.bac || id: bac, nolog

Mixed-effects ML regression                     Number of obs      =       204
Group variable: id                              Number of groups   =        51

                                                Obs per group: min =         4
                                                               avg =       4.0
                                                               max =         4


                                                Wald chi2(3)       =     56.95
Log likelihood =  -170.9156                     Prob > chi2        =    0.0000

------------------------------------------------------------------------------
     ratefem |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.gender |  -.6280836   .3203292    -1.96   0.050    -1.255917   -.0002499
         bac |   3.433733   .5231428     6.56   0.000     2.408392    4.459074
             |
gender#c.bac |
          1  |  -1.154182   .9270306    -1.25   0.213    -2.971128    .6627648
             |
       _cons |   6.442059   .1903235    33.85   0.000     6.069032    6.815086
------------------------------------------------------------------------------
```

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
id: Independent				
sd(bac)	1.621421	.6564771	.7332693	3.585323
sd(_cons)	1.056773	.1087889	.8636849	1.293029
sd(Residual)	.3371602	.02408	.2931186	.387819

```
LR test vs. linear regression:      chi2(2) =    279.98   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

هذه النتائج لصور الإناث تدعم فرضية أكواب الجُعّة: فمعدلات الجاذبية لوجوه الإناث تزداد بزيادة نسبة الكحول في الدم، الجنس له تأثير لا يكاد أن يكون ملحوظاً: فالنساء أعطين معدلات أقل بقليل لوجوه النساء، تفاعل المتغيرين *gender×bac* ليس ذا معنوية إحصائية. وعموماً فإن الميل العشوائي والتقاطع العشوائي للمتغير *bac* شهد تبايناً ذا معنوية إحصائية، وتضمن اختلافات جوهرية من شخص لآخر في متوسط المعدلات المعطاة، وكيفية تأثير الكحول على هذه المعدلات.

الأمر **margins** والأمر **marginsplot** يساعدان في عرض هذه النتائج بيانياً، فمخرجات هذين الأمرين لم تُعرض هنا، ولكن سوف يتم دمجها مع التحليل التالي لتشكل الشكل (8.13).

```
.margins, at(bac= (0(.2).4) gender=(0 1)) vsquish
.marginsplot, title("Female photos") ytitle("")
 xtitle("") noci
 legend(position(11) ring(0) row(2)
 title("Gender", size(medsmall)))
 ylabel(4(1)8, grid gmin gmax)
 plot1opts(lpattern(solid) msymbol(T))
 plot2opts(lpattern(dash) msymbol(Oh)) saving
 (fig13_08RF)
```

الأمر **xtmixed** الثاني يقوم بصياغة نموذج لمعدلات صور الذكور (*ratemal*):

```
.xtmixed ratemal i.gender##c.bac || id: bac,
 nolog
```

```
Mixed-effects ML regression                     Number of obs      =       201
Group variable: id                              Number of groups   =        51

                                                Obs per group: min =         3
                                                               avg =       3.9
                                                               max =         4

                                                Wald chi2(3)       =     32.74
Log likelihood = -221.83425                     Prob > chi2        =    0.0000

------------------------------------------------------------------------------
    ratemale |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.gender |   2.011293   .3742453     5.37   0.000     1.277786    2.744801
         bac |   .6401159   .7601351     0.84   0.400    -.8497215    2.129953
             |
gender#c.bac |
           1 |   .6055665   1.328251     0.46   0.648    -1.997758    3.208891
             |
       _cons |   3.946884   .2224468    17.74   0.000     3.510897    4.382872
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
id: Independent              |
                     sd(bac) |   2.738856    .535468      1.867035    4.017777
                   sd(_cons) |   1.223205   .1284553      .9956575    1.502756
-----------------------------+------------------------------------------------
                sd(Residual) |   .4408696   .0278099       .389598    .4988886
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(2) =   255.98   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

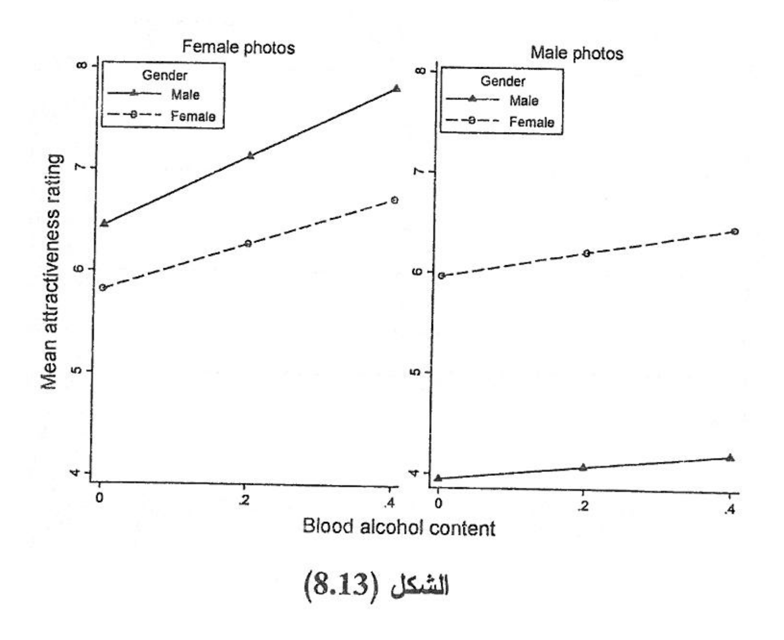
لا يبدو أن هناك تأثيرًا لأكواب الجعة على كيفية تقييم صور الـــذكور. ومن ناحية أخرى، فإن الجنس كان له تأثير ثابت أقوى، أما تفاعل المتغيرين *gender, bac* لم يكن ذا معنوية إحصائية، ولكننا مرة أخرى رأينـــا تباينـــاً ذا معنوية في الميل والتقاطع العشوئي.

```
.margins, at(bac = (0(.2).4) gender = (0 1))
 vsquish
.marginsplot, title("Male photos") ytitle("")
 xtitle("") noci
 legend(position(11) ring(0) row(2)
 title("Gender", size(medsmall)))
 ylabel(4(1)8, grid gmin gmax)
 plot1opts(lpattern(solid) msymbol(T))
```

```
plot2opts(lpattern(dash) msymbol(Oh))
saving(fig13_08RM)
```

الشكل (8.13) يدمج الشكلين البيـــانيين marginsplot وذلـــك للمقارنـــة المباشرة.

```
.graph combine fig13_08RF.gph fig13_08RM.gph,
imargin(vsmall) l1("Mean attractiveness
rating")b2("Blood alcohol content")
```



الشكل (8.13)

الجانب الأيسر من الشكل (8.13) يعرض التأثير الجـــوهري للكحـــول على تقييم صور الإناث. المشاركون من الذكور قاموا بإعطاء معدلات أعلى نوعاً ما وظهر بأنهم كانوا أكثر تأثراً بالكحول. الجانب الأيمن يصوّر وضعاً مختلفاً، لتقييم صور الذكور. كما أن المشاركات الإناث قمن بإعطاء معدلات أعلى بشكل ملحوظ، ولكن تقييم المشاركين من الذكور والإناث لـــم يتغيـــر بصورة كبيرة كلما زاد تناول الكحول.

السلاسل الزمنية المقطعية : Cross-Sectional Time Series

هذا الجزء يقوم بتطبيق الأمر xtmixed على نوع مختلف من البيانـــات متعددة المستويات، وهو الـــسلاسل الزمنيـــة المقطعيـــة، ملـــف البيانـــات

Alaska_regions.dta يحتوي على سلاسل زمنية للسكان في 27 قرية أو بلدية أو مناطق تعداد والتي تمثل معاً ولاية آلاسكا، هذه 27 منطقة هي جزء من إطار لقاعدة بيانات سمات الإنسان لعموم القطب الشمالي تم شرحها بواسطة Hamilton and Lammers (2011)، في بيانات الملف *Alaska_regions.dta* هناك متغير وهمي اسمه *large* يمثل أكثر خمس مناطق ازدحاماً بالسكان، والتي كان بها عدد السكان في سنة 2011 أكبر من 20,000، المناطق الأخرى الـ 22 يغلب عليها الطابع الريفي وسكانها أقل وربما متوزعون على نطاق واسع. فمثلاً القرية بشمال غرب القطب الشمالي تغطي منطقة جغرافية أكبر من ولاية Maine بالولايات المتحدة، ولكن عدد سكانها أقل من 8,000 نسمة، لكل منطقة من المناطق 27 هناك بيانات لعدد من السنوات تتراوح من سنة 1969 إلى 2011 ولكن مع وجود العديد من القيم المفقودة، ولذا فإنه لكل متغير هناك 27 نظيرًا ولكنه في العادة عبارة عن سلاسل زمنية ناقصة.

```
.use C:\data\Alaska_regions.dta, clear
.describe

Contains data from C:\data\Alaska_regions.dta
  obs:           852                          Alaska regions population 1969-2011
 vars:             7                          2 Jul 2012 06:11
 size:        44,304
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
regionname      str34  %34s                   Region name
regioncode      float  %9.0g                  AON-SI region code
year            int    %9.0g                  Year
pop             double %12.0g                 Population in thousands
large           byte   %9.0g       large      Regions 2011 population > 20,000
year0           byte   %9.0g                  years since 1968
year2           int    %9.0g                  years0 squared
-------------------------------------------------------------------------------
Sorted by:  regionname  year
```

خلال النصف الأول من الفترة الزمنية التي تغطيها البيانات، لوحظ أن عدد السكان ازداد بصورة كبيرة في العديد من المناطق الريفية في آلاسكا. وعموماً فإنه في السنوات الأخيرة، فإن معدل النمو انخفض، وفي بعض المناطق انخفض عدد السكان. هذه الاتجاهات لها علاقة بالجدل الدائر حول

النمو الاقتصادي المستدام لهذه المناطق، وأيضاً الأهمية الثقافية لسكان آلاسكا الأصليين الذين يعيشون هناك.

ولأن اتجاهات عدد السكان ببساطة لم تزد، فلا يمكننا صياغة نموذج واقعي كدالة خطية للسنوات *year*. فالنموذج المختلط أدناه يمثل اتجاه عدد السكان كدالة تربيعية تقوم بحساب انحدار عدد السكان بالآلاف (*pop*) على السنوات منذ سنة 1968 (*year0*) وأيضاً على تربيع *year0* ، سوف نسمح للتقاطع الثابت (β) والعشوائي (u_j) وقيم الميل لكلا الحدين، أكبر خمس مناطق مزدحمة بالسكان تم استبعادها في هذا التحليل حتى نركز على المناطق الريفية بآلاسكا.

$$\text{عدد السكان } population_{ij} = \beta_0 + \beta_1 year0_{ij} + \beta_2 year0^2_{ij} + u_{0j} + u_{1j} year0_{ij} + u_{2j} year0^2_{ij} + \in_{ij} \quad [7.13]$$

```
.keep if large == 0
.xtmixed pop year0 year2 ||  regionname: year0
 year2
```

```
Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log likelihood = -457.61229
Iteration 1:   log likelihood = -457.61196
Iteration 2:   log likelihood = -457.61196

Computing standard errors:

Mixed-effects ML regression                     Number of obs      =       639
Group variable: regionname                      Number of groups   =        22

                                                Obs per group: min =        22
                                                               avg =      29.0
                                                               max =        40


                                                Wald chi2(2)       =      1.22
Log likelihood = -457.61196                     Prob > chi2        =    0.5424
```

pop	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
year0	.0615239	.0783008	0.79	0.432	-.0919428	.2149906
year2	-.0008945	.0010787	-0.83	0.407	-.0030087	.0012197
_cons	5.457939	1.342116	4.07	0.000	2.82744	8.088438

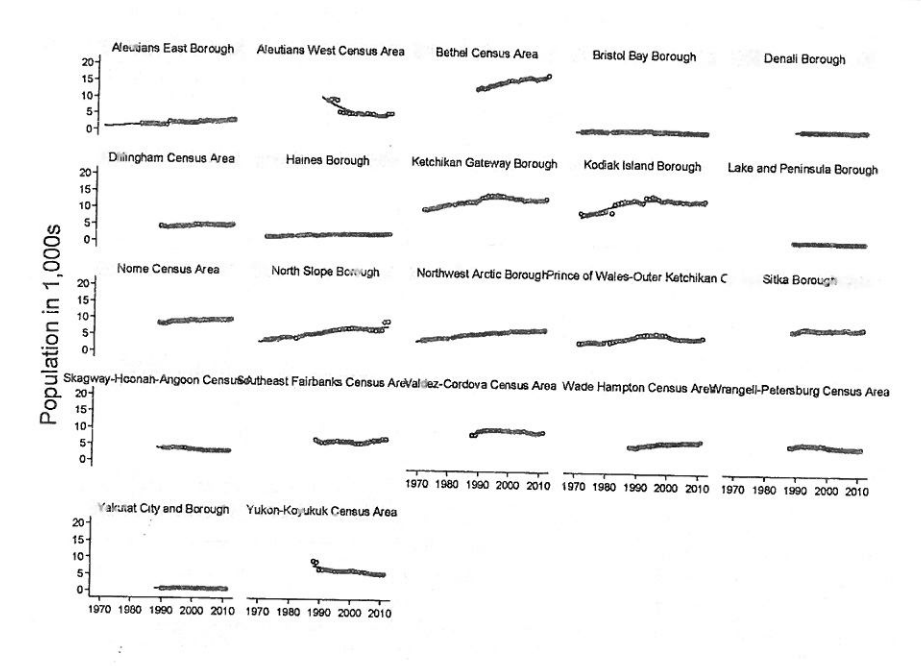
Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
regionname: Independent				
sd(year0)	.3563579	.0619934	.2534009	.5011464
sd(year2)	.004861	.0008663	.0034278	.0068933
sd(_cons)	6.145796	1.04392	4.405485	8.573587
sd(Residual)	.3524305	.0108473	.3317988	.3743451

```
LR test vs. linear regression:        chi2(3) =  2703.76   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

كل التأثيرات العشوائية توضح تبايناً ذا معنوية إحصائية من مكان لآخر. ومن ناحية أخرى، فإن مُعاملات التأثير الثابت للمتغير *year0* والمتغير *year2* لا تختلف بدرجة كبيرة عن الصفر، مشيرةً إلى النقص في وجود نمط عام لـ 22 منطقة، الرسم البياني يتنبأ بعدد السكان (مع منحنى وسيط محدب) مع عدد السكان الفعلي والسنة، وهذا يساعد في عرض تفاصيل التباين من منطقة لأخرى، ويشرح عدم قدرة الأمر **xtmixed** على إيجاد اتجاه عام (الشكل 9.13)، وفي بعض المناطق فإن عدد السكان زاد بثبات، بينما في مناطق أخرى فإن اتجاه النمو انخفض أو انعكس. النموذج يقوم بعمل جيد لتقليل بعض الفجوات الواضحة في البيانات مثل استبعاد عدد السكان في جُزر اليوتانيس الغربية Aleutians West التي شهدت انخفاضاً بعد تقليص المحطة الجوية البحرية في سنة 1994 أو الزيادة في المنحدر الشمالي North Slope التي حدثت بعد زيادة العاملين في المناطق البعيدة وكانت هذه الزيادة واضحة في تعداد سنة 2010.

```
.predict yhat, fitted
.graph twoway scatter pop year, msymbol(Oh)
 ||    mspline   yhat   year,   lwidth(medthick)
 bands(50)
 || , by(regionname, note("") legend(off))
 ylabel(0(5)20, angle(horizontal)) xtitle("")
 ytitle("Population          in         1,000s")
 xlabel(1970(10)2010, grid)
```



الشكل (9.13)

سوف نقوم بتحليل أكثر موضوعية، وذلك باستخدام انحدار التـأثيرات المختلطة لصياغة نموذج للعلاقات يتضمن سلاسل زمنية متعددة يظهر في ورقة بحثية حول تعداد السكان والمناخ والكهرباء المستخدمة في مدن وقرى المنطقة القطبية الشمالية في آلاسكا (2011 Hamilton et al.)، بيانات الملف *Alaska_places.dta* تحتوي على البيانات الأساسية لهذا التحليل، وتتضمن سلاسل زمنية سنوية لكل من 42 قرية ومدينة في المنطقة القطبية الشمالية في آلاسكا وكلها ضمن مناطق التعداد الخمس أو القرى التي تم تمثيلها بيانياً في الشكل (9.13) أعلاه. المتغيرات تتضمن السكان وكيلووات ساعة للكهرباء المباعة والمعدل المتوسط لتكلفة الكهرباء (التي كانت سائدة في سنة 2009 بالدولار) وتقديرات درجات حرارة الصيف، ومعدل الهطول حول تلك المنطقة في كل سنة، الورقة تزودنا بمعلومات حول تعريفات المتغيرات، ومصادر البيانات، وشرح لهذا التحليل.

```
.use C:\data\Alaska_places.dta, clear
.describe
```

```
Contains data from C:\data\Alaska_places.dta
  obs:           742                          Population, climate & electricity use in the Arctic
                                                (Hamilton 2011)
 vars:            12                          2 Jul 2012 06:11
 size:        57,876

              storage  display     value
variable name   type   format      label      variable label

regionname      str24  %24s                   Region name
regioncode      long   %9.0g                  AON-SI region code
place           str21  %21s                   Place name
placecode       byte   %21.0g      placecode
                                              Place code (labeled)
year            int    %ty                    Year
pop             int    %12.0g                 Population--est. Jul 1/Census Apr 1
logpop          float  %9.0g                  log10(pop)
kwhsold20       float  %9.0g                  kWh sold ajusted if 9-11 months, millions
logkwhsold      float  %9.0g                  log10(kwhsold20)
rateres09       float  %9.0g                  av. res. rate 2009$ = rateres*cpianc09
fsumtempD       float  %9.0g                  UDel FY summer (L1.Jul-Sep & May-Jun) temp
fsumprecD       float  %9.0g                  UDel FY summer (L1.Jul-Sep & May-Jun) prec

Sorted by:  placecode  year
```

هذه البيانات تم اعتبارها بيانات طولية عن طريق الأمر **xtset**، والـــذي قام بتحديد المتغير *placecode* كمتغير طولي، والسنة *year* كمتغير زمني:

```
.xtset placecode year

 panel variable:  placecode (unbalanced)
  time variable:  year, 1990 to 2008, but with gaps
          delta:  1 year
```

تحليل التأثير المختلط أدناه يقوم بصياغة نمـــوذج للكيلـــووات ســـاعة للكهرباء المباعة في كل حي وكل سنة كدالة لعدد السكان، ومعدل الكهربـــاء بالدولار وفقاً لسنة 2009 وحرارة الصيف والهطـــول والـــسنة. النمـــوذج يتضمن تأثيرات عشوائية لكل من التعداد السكاني، وأخطاء الانحدار الـــذاتي من الدرجة الأولى. السبب وراء صياغة النموذج بهذا الشكل واختبارات الثقة لنتائج النموذج تم عرضها في الورقة. الاختصار في العرض هنا يهدف إلى توضيح ما سوف يكون عليه مثل هذا النوع من التحليل.

```
.xtmixed logkwhsold logpop rateres09 fsumtempD
  fsumprecD year
```

```
|| placecode: logpop, nocons residual(ar 1,
t(year))nolog reml

Note: time gaps exist in the estimation data

Mixed-effects REML regression                   Number of obs      =       742
Group variable: placecode                       Number of groups   =        42

                                                Obs per group: min =        12
                                                               avg =      17.7
                                                               max =        19

                                                Wald chi2(5)       =    359.12
Log restricted-likelihood =   1008.232          Prob > chi2        =    0.0000
```

logkwhsold	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
logpop	.7086409	.0716509	9.89	0.000	.5682078	.849074
rateres09	-.0011494	.0005259	-2.19	0.029	-.0021801	-.0001187
fsumtempD	-.0038939	.0018784	-2.07	0.038	-.0075755	-.0002123
fsumprecD	.000272	.0001416	1.92	0.055	-5.57e-06	.0005495
year	.012952	.0010914	11.87	0.000	.0108129	.0150911
_cons	-27.51866	2.153197	-12.78	0.000	-31.73885	-23.29847

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
placecode: Identity				
sd(logpop)	.0989276	.0132542	.0760807	.1286354
Residual: AR(1)				
rho	.7900083	.0394952	.699089	.8557882
sd(e)	.0857878	.0076267	.0720696	.1021173

```
LR test vs. linear regression:       chi2(2) =   1506.36   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

النموذج هو:

$$\log(kwhres_{it}) = -27.52 + 0.70861\log_{10}(pop_{it}) - 0.0011rateres09_{it} - 0.0039fsumtep_{it} + 0.0003fsumprec_{it} + 0.130t + \mu_i\log_{10}(pop_{it}) + 0.7900\square_{i,t-1} + u_{it} \quad [13.8]$$

استخدام الكهرباء في قُرى المنطقة القطبية الشمالية تم التنبؤ به بواسطة عدد السكان، والسعر، ودرجة حرارة الصيف. وعلى عكس الوضع في المناطق الجنوبية، حيث يكون فصل الصيف حاراً والذي يعني استخداماً أكثر للكهرباء بسبب المكيفات، فإنه إذا كان صيف المناطق القطبية الشمالية حاراً

(وفي العادة أقل مطراً) فقد يؤدي ذلك إلى تشجيع الناس إلى قضاء أوقات أكبر خارج بيوتهم، وبتعديل النموذج للأخذ في الاعتبار تـأثيرات عـدد الـسكان والسعر والجو، فإننا نرى نمطاً تصاعدياً عاماً في اسـتخدام الكهربـاء، لأن الحياة تصبح أكثر تطلباً للكهرباء. وأخيراً فإن التباين الكبير في الميل العشوائي لعدد السكان يشير إلى أن تأثيرات كل شخص تختلف من مكانٍ لآخر، وهـذه التأثيرات تميل لتكون أكبر في المناطق الشمالية ومجتمعات مناطق المنحـدر الشمالية، وحد الخطأ للانحدار الذاتي AR(1) يظهر أيضاً ذو معنوية إحصائية.

وكما لاحظنا في الفصل (12) فإن السلاسل الزمنية عادة يـتم اختبـار استقرارها stationarity قبل صياغة النموذج، وذلك لتجنب الحـصول علـى نتائج مضللة، الفروقات تعتبر إحدى الأدوات المفيدة للقيام بذلك. وبدلاً مـن ذلك، فإنه حتى عندما تكون السلاسل الزمنية غير مستقرة – كما فـي هـذا المثال (أو كما في نماذج ARMAX في الفصل 12) – فإنه يمكننا البحث عن مزيج خطّي من المتغيرات المستقرة (لمزيد من المعلومـات عـن التكامـل المشترك cointegration انظر كتاب Hamilton 1994) النموذج [8.13] يمكنـه القيام بهذه المهمة بكفاءة عالية، كل 42 سلسلة المتبقية (سلسلة واحـدة لكـل مجتمع) والتي تم إنشاؤها بواسطة هذا النموذج لم يُظهر أي منهـا ارتباطـاً ذاتياً ذا معنوية إحصائية في اختبار إحصائيات Q الصندوقية. ولـذلك فـإن البواقي لا يمكن تمييزها عن قيم الضجيج الأبيض، وقـيم عمليـة التغـاير الثابت. الأوامر أدناه تقوم بحساب القيم المتوقعة آخذةً فـي الاعتبـار حـد الانحدار الذاتي (*yhat_xt*) ثم يتم استخدام هذه التوقعـات لحـساب البـواقي (*resid-xt*)، اختبارات Q للضجيج الأبيض (wntestq) لا تعرض انحدارًا ذاتياً متبقياً. والمخرجات أدناه هي لأول 3 من 42 مجتمعاً.

```
.predict yhat_xt, fitted
.generate resid_xt = logkwhsold - yhat_xt
.replace yhat_xt= yhat_xt +(.7900077*L1.resid_xt)
.gen yhat_xt10 = 10^yhat_xt
.replace resid_xt = logkwhsold - yhat_xt
.label variable yhat_xt "predicted values
log(million kWh)"
.label variable yhat_xt10 "predicted values in
 millions of kWh"
```

```
.label variable resid_xt "residuals log(million kWh)"
.wntestq resid_xt if place == "Ambler city", lags(5)

 Portmanteau test for white noise
 ---------------------------------------
  Portmanteau (Q) statistic =     4.3048
  Prob > chi2(5)            =     0.5064

.wntestq resid_xt if place == "Anaktuvuk Pass city", lags(5)
 Portmanteau test for white noise
 ---------------------------------------
  Portmanteau (Q) statistic =     2.3503
  Prob > chi2(5)            =     0.7989

.wntestq resid_xt if place == "Aniak city", lags(5)
 Portmanteau test for white noise
 ---------------------------------------
  Portmanteau (Q) statistic =     5.6826
  Prob > chi2(5)            =     0.3383
```

اختبارات مشابهة لكل 42 مجتمعاً وجدت بأن جميع سلاسل البواقي ليس لها ارتباط ذاتي ذو معنوية.

الانحدار اللوغاريتمي ذو التأثيرات المختلطة :
Mixed-Effects Logit Regression

منذ سنة 1972 يحاول الاستطلاع الاجتماعي العام (Davis et al. 2005) بمتابعة الرأي العام الأمريكي من خلال سلاسل استطلاع سنوية أو نصف سنوية. وهذه البيانات متوافرة للتدريس والبحث، ملف البيانات *GSS_2010_SwS* يحتوي على عينة فرعية صغيرة لمتغيرات ومشاهدات من استطلاع سنة 2010، وهي تتضمن متغيرات إضافية مع إجابات عن أسئلة حول التصويت والمخدرات والرقابة على السلاح والتغير المناخي والتطور؛ موقع الاستطلاع الاجتماعي العامGSS على الإنترنت يوفر معلومات مفصلة عن مصدر هذه البيانات (http://www3.norc.org/GSS+Website).

```
.use C:\data\GSS_2010_SwS.dta, clear
.describe
```

```
Contains data from C:\data\GSS_2010_SwS.dta
  obs:           809                          General Social Survey 2010--evolution etc.
 vars:            19                          2 Jul 2012 06:11
 size:        21,843
--------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
--------------------------------------------------------------------------------
id              int    %8.0g                  Respondent ID number
year            int    %8.0g                  GSS year
wtssall         float  %9.0g       LABCM      probability weight
cendiv          byte   %15.0g      cendiv     Census division
logsize         float  %9.0g                  log10(size place in 1,000s, +1)
age             byte   %8.0g       age        Age in years
nonwhite        byte   %9.0g       nonwhite   Consider self white/nonwhite
sex             byte   %8.0g       sex        Respondent gender
educ            byte   %8.0g       educ       Highest year of schooling
married         byte   %9.0g       yesno      Currently married
income06        byte   %15.0g      income     Total family income
polviews        byte   %12.0g      polviews   Polit views liberal-conservative
bush            byte   %9.0g       yesno      Voted for Bush in 2004
obama           byte   %9.0g       yesno      Voted for Obama in 2004
postlife        byte   %8.0g       yesno      Believe in life after death
grass           byte   %9.0g       grass      Should marijuana be legalized?
gunlaw          byte   %9.0g       gunlaw     Oppose permit required to buy gun
sealevel        byte   %10.0g      sealevel   Bothered if sea level rose 20 ft
evolve          byte   %9.0g       true       Humans developed/ earlier species
--------------------------------------------------------------------------------
Sorted by:  id
```

سؤال GSS حول التطور سوف يكون المحور الأساسي لهذا التحليل. هذا السؤال يحتوي على جزء من القدرة على القراءة والكتابة والعلوم ويسأل عما إذا كانت العبارة أدناه صحيحة أو خاطئة.

البشر – كما نعرفهم اليوم – تطوّروا من سُلالات سابقة من الحيوانات.

هذا السؤال يستند إلى اعتقادات فردية، بالإضافة إلى معرفـــة علميـــة، حوالي 55% من المشاركين قالوا إن هذه العبارة صحيحة.

```
.tab evolve
```

```
     Humans |
  developed/ |
     earlier |
     species |      Freq.     Percent        Cum.
-------------+-----------------------------------
       False |        360       44.50       44.50
        True |        449       55.50      100.00
-------------+-----------------------------------
       Total |        809      100.00
```

لا توجد قيم مفقودة للمتغير *evolve* ولا يوجد معيار تم وضعه لاختيار هذا الجزء الفرعي من البيانات والذي يمثل 809 مشاركين في استطلاع GSS، الإجابات التي تقول بأن العبارة "خاطئة" “False” في المتغير *evolve* تم ترميزها بالرقم 0 والعبارة "صحيحة" “True” 1، السؤال حول الاستطلاع المرجح يُعتبر مسألة معقدة مع النماذج متعددة المستويات. وهذه المسألة لم يتم حلها بعد في أمر ستاتا **xtmelogit** ولن يتم تناولها هنا.

وفي العادة، فإن البحوث تجد أن زيادة القدرة على القراءة والكتابة والعلوم تزداد مع التعليم، وأيضاً تتعلق بعوامل أخرى تتعلق بخلفية الشخص، وفي حالة المتغير *evolve* فإننا نتوقع أن هذا المتغير له علاقة ما مع التوقعات السياسية أيضاً، فالانحدار اللوغاريتمي البسيط يؤكد مثل هذه الفرضيات، حيث إن النتائج تُشير إلى أن الذكور ذوي المستوى التعليمي وذوي وجهات سياسية معتدلة في الغالب يعتقدون بتطور الإنسان من سلالات حيوانية سابقة.

```
.logit evolve sex age educ polviews, nolog

Logistic regression                     Number of obs   =        785
                                        LR chi2(4)      =      98.93
                                        Prob > chi2     =     0.0000
Log likelihood = -489.36806             Pseudo R2       =     0.0918
```

evolve	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sex	-.6089296	.1565972	-3.89	0.000	-.9158545	-.3020047
age	-.008189	.0045313	-1.81	0.071	-.0170701	.0006922
educ	.0990929	.0254359	3.90	0.000	.0492395	.1489463
polviews	-.4482161	.0575731	-7.79	0.000	-.5610573	-.3353749
_cons	1.457699	.4891102	2.98	0.003	.4990611	2.416338

وبغض النظر عن مثل هذه المتغيرات التنبؤية على المستوى الفردي بخصوص الاعتقاد حول التطور البشري، فإنه قد تكون هناك مكونات على مستوى المناطق أيضاً، الخلافات حول تدريس التطور البشري في المدارس كان السمة البارزة في الجنوب. واتساقاً مع هذا الانطباع، فإن اختبار كاي

تربيع لبيانات GSS يوضح فروقات ذات معنوية إحصائية بين تقسيمات تعداد السكان بالولايات المتحدة، قبول الفرضية مرتفع جداً (89%) بين المشاركين في الدراسة من نيوإنجلاند New England وهي الولايات Maine, Massachusetts, New Hampshire, Rhode Island, Vermont، وهذه النسبة تكون عند أقل قيمة لها (36%) بين تقسيم السكان للمناطق الوسطى الجنوبية الشرقية E South Central وهي الولايات Alabama, Kentucky, Mississipi, Tennessee، تقسيمات السكان في مناطق المحيط الهادئ Pacific والمناطق الوسطى الجنوبية الغربية W South Central و هي ثاني أعلى وثاني أقل قيم على التوالي.

```
.tab cendiv evolve, row nof chi2
```

Census division	Humans developed/ earlier species False	True	Total
New England	11.11	88.89	100.00
Middle Atlantic	39.00	61.00	100.00
E North Central	43.88	56.12	100.00
W North Central	42.00	58.00	100.00
South Atlantic	50.81	49.19	100.00
E South Central	64.29	35.71	100.00
W South Central	62.34	37.66	100.00
Mountain	43.55	56.45	100.00
Pacific	27.43	72.57	100.00
Total	44.50	55.50	100.00

Pearson chi2(8) = 48.6890 Pr = 0.000

يمكننا إضافة تقسيم تعداد سكان لتحليل الانحدار كمجموعة متغيرات إشارية للتقسيمات من 2 وحتى 9 كل منها يتناقض مع تقسيم السكان 1 الخاص بنيوانجلاند New England.

```
.logit evolve sex age educ polviews i.cendiv,
 nolog
```

```
Logistic regression                           Number of obs   =        785
                                              LR chi2(12)     =     124.92
                                              Prob > chi2     =     0.0000
Log likelihood = -476.37206                   Pseudo R2       =     0.1159
```

evolve	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sex	-.5609946	.160387	-3.50	0.000	-.8753473	-.2466419
age	-.0092908	.0046327	-2.01	0.045	-.0183706	-.0002109
educ	.0842967	.0261043	3.23	0.001	.0331333	.1354601
polviews	-.416007	.0591817	-7.03	0.000	-.532001	-.3000131
cendiv						
2	-1.501592	.6612973	-2.27	0.023	-2.797711	-.2054736
3	-1.602085	.6504787	-2.46	0.014	-2.877	-.3271704
4	-1.505793	.6931599	-2.17	0.030	-2.864361	-.1472243
5	-1.843963	.6442829	-2.86	0.004	-3.106734	-.5811918
6	-2.149803	.6973044	-3.08	0.002	-3.516495	-.7831115
7	-2.239585	.6743959	-3.32	0.001	-3.561376	-.9177932
8	-1.454279	.6854426	-2.12	0.034	-2.797722	-.1108363
9	-1.141026	.6642829	-1.72	0.086	-2.442996	.1609447
_cons	3.179554	.8138406	3.91	0.000	1.584455	4.774652

المُعاملات الخاص بالمتغيرات الإشارية لتقسيم ما تعطي تغييرًا في التقاطع $-y$ لذلك التقسيم بالمقارنة مع نيوإنجلاند، كل هذه المُعاملات سالبة، لأن الاحتمالات اللوغاريتمية لقبول الاعتقاد بالتطور البشري أقل في تقسيمات السكان الأخرى عنه في نيوإنجلاند. وكما هو متوقع، فإن الفرق كبير جداً للتقسيم 6 الذي يمثل المنطقة الوسطى الجنوبية الشرقية E South Central، أما التقسيم 9 الذي يمثل منطقة المحيط الهادئ Pacific هو الوحيد الذي لا يختلف كثيراً واختلافه ليس ذا معنوية عن نيوإنجلاند، صافي هذه التأثيرات على مستوى المناطق وكل المتغيرات التنبؤية على المستوى الفردي أظهرت تأثيرات ذات معنوية في الاتجاهات المتوقعة.

طريقة المتغير الإشاري تعمل بشكل جيد هنا، وذلك لأن لدينا 9 مجموعات فقط (تقسيمات سكان)، وهذه التقسيمات تختبر فرضيات بسيطة حول حركة التقاطعات $-y$، وعند وجود مجموعات أكثر أو فرضيات معقدة فإن طريقة التأثيرات المختلطة يمكن أن تكون أكثر عملية، فمثلاً قد نقوم

بإدخال تقاطعات عشوائية لكل تقسيم في التعداد السكاني في نموذج انحـــدار لوغاريتمي ذي تأثيرات مختلطة كما سيتم لاحقاً، تركيبة الأمـــر **xtmelogit** تتشابه مع تلك الخاصة بالأمر **xtmixed**.

```
.xtmelogit evolve sex age educ polviews ||
 cendiv: , nolog

Mixed-effects logistic regression               Number of obs      =       785
Group variable: cendiv                          Number of groups   =         9

                                                Obs per group: min =        27
                                                               avg =      87.2
                                                               max =       181

Integration points =   7                        Wald chi2(4)       =     72.41
Log likelihood = -487.10546                     Prob > chi2        =    0.0000

------------------------------------------------------------------------------
      evolve |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |  -.5794058   .1591076    -3.64   0.000    -.8912511   -.2675606
         age |  -.0086106   .0045962    -1.87   0.061    -.0176191    .0003979
        educ |   .0910441   .0259804     3.50   0.000     .0401235    .1419647
    polviews |  -.4300722   .0588037    -7.31   0.000    -.5453254   -.3148191
       _cons |   1.541323   .5135582     3.00   0.003     .5347679    2.547879
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
cendiv: Identity             |
                   sd(_cons) |   .3375876   .1559346      .1365241    .8347641
------------------------------------------------------------------------------
LR test vs. logistic regression: chibar2(01) =     4.53 Prob>=chibar2 = 0.0167
```

التقاطعات العشوائية في المخرجات أعلاه توضح تبايناً ذا معنوية، وهذا ما يوضحه اختبار معدل الأرجحية العظمى مع الانحدار اللوغاريتمي العادي ($p = 0.0167$) أو من خلال الانحـــراف المعيـــاري للتقاطعـــات العـــشوائية (0.3376)، حيث إنه أكبر من ضعف خطأه المعياري (0.1559)، الأوامـــر التالية تقوم بتقدير قيم لهذه التقاطعات العشوائية من خلال الأمر **predict** ثـــم إنشاء جدول يعرض هذه القيم بواســـطة المتغيـــر *cendiv*، وبالاتـــساق مـــع تحليلات كاي تربيع السابقة، وتحليلات المتغيرات الإشارية الـــسابقة، فإننـــا نرى تقاطعات y– عشوائية موجبة (مؤدياً إلى زيادة التأثير الكلي) لتقسيمات

السكان في نيوإنجلاند والمحيط الهادئ، ولكن هناك تقاطعات y- عشوائية سالبة (مؤدياً إلى انخفاض التأثير الكلي) لتقسيمات السكان بالمنطقة الوسطى الجنوبية الشرقية والمنطقة الوسطى الجنوبية الغربية.

```
.predict recendiv, reffects
.label variable recendiv "random-effect
 intercept cendiv"
.table cendiv, contents(mean recendiv)
```

Census division	mean(recendiv)
New England	.4649539
Middle Atlantic	.0523787
E North Central	-.0165851
W North Central	.0429461
South Atlantic	-.2134227
E South Central	-.3085577
W South Central	-.4224425
Mountain	.083739
Pacific	.3052153

ولذا فعند استخدام أي طريقة سوف نجد نمطاً موثوقاً للاختلافات الإقليمية حول الاعتقاد بتطور الإنسان، حتى بعد التحكم في العوامل الفردية. فنماذج التأثيرات المختلطة تسمح لنا بالتقدم أكثر من خلال اختبار أفكار أكثر تفصيلاً حول الاختلافات الإقليمية.

بعض الدراسات حددت التعليم كمؤثر أساسي في الاعتقاد بتطور الإنسان، وبعض الاعتقادات العلمية الأخرى، والسؤال الآن هو: هل تأثيرات التعليم تختلف من تقسيم سكاني لآخر؟ يمكننا اختبار ذلك بواسطة إضافة قيم ميل وتقاطعات عشوائية معاً:

```
.xtmelogit evolve sex age educ polviews ||
 cendiv: educ, nolog
```

```
Mixed-effects logistic regression               Number of obs      =       785
Group variable: cendiv                          Number of groups   =         9

                                                Obs per group: min =        27
                                                               avg =      87.2
                                                               max =       181

Integration points =   7                        Wald chi2(4)       =     71.63
Log likelihood = -486.57368                     Prob > chi2        =    0.0000
```

evolve	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sex	-.5692675	.1595327	-3.57	0.000	-.8819458	-.2565893
age	-.0090823	.0046088	-1.97	0.049	-.0181153	-.0000492
educ	.0924205	.027522	3.36	0.001	.0384784	.1463627
polviews	-.4290164	.0588184	-7.29	0.000	-.5442984	-.3137343
_cons	1.532699	.4979934	3.08	0.002	.55665	2.508748

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
cendiv: Independent				
sd(educ)	.0268375	.0115911	.0115109	.0625712
sd(_cons)	5.23e-07	.4785104	0	.

```
LR test vs. logistic regression:     chi2(2) =     5.59   Prob > chi2 = 0.0612

Note: LR test is conservative and provided only for reference.
```

الانحراف المعياري لقيم الميل العشوائي للتعليم أكثر من ضعف الخطأ المعياري، وهذا يشير إلى أن التباين الإقليمي ذو معنوية إحصائية، ومن ناحية أخرى، فإن الانحراف المعياري للتقاطعات العشوائية يقترب من الصفر، وهذا يعني عدم وجود تباين من مكان لآخر، قد يتم استخدام نموذج أبسط يقوم بإهمال التقاطعات العشوائية من خلال الخيار **nocons** الذي يعطي أرجحية لوغاريتمية متطابقة.

```
.xtmelogit evolve sex age educ polviews
  || cendiv: educ, nolog nocons
```

```
Mixed-effects logistic regression               Number of obs      =       785
Group variable: cendiv                          Number of groups   =         9

                                                Obs per group: min =        27
                                                               avg =      87.2
                                                               max =       181

Integration points =   7                        Wald chi2(4)       =     71.63
Log likelihood = -486.57368                     Prob > chi2        =    0.0000
```

evolve	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sex	-.5692676	.1595327	-3.57	0.000	-.8819459	-.2565893
age	-.0090823	.0046088	-1.97	0.049	-.0181153	-.0000492
educ	.0924205	.027522	3.36	0.001	.0384784	.1463626
polviews	-.4290164	.0588184	-7.29	0.000	-.5442984	-.3137343
_cons	1.532699	.4979933	3.08	0.002	.55665	2.508748

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
cendiv: Identity				
sd(educ)	.0268374	.011591	.0115108	.062571

LR test vs. logistic regression: chibar2(01) =   5.59 Prob>=chibar2 = 0.0090

نموذج الميل العشوائي المبسط أعلاه، يقوم على افتراض أن التعليم له تأثيرات مختلفة على الاعتقاد بتطور الإنسان في أجزاء مختلفة من الدولة، ولمعرفة ماهي هذه التأثيرات، يمكننا توقع **predict** قيم الميل العشوائي، وإنشاء متغير جديد باسم *raneduc*، التأثيرات الكلية للمتغير *educ* تساوي هذه التأثيرات العشوائية زائد التأثير الثابت لمعامل _b[*educ*]، ثابت "المتغير" المُسمى *fixeduc* يتم إنشاؤه لعرض التأثيرات الثابتة في جدول ومتغير جديد يُسمى *toteduc* يعرض التأثيرات الكلية للتعليم أو الميل على *educ* لكل تقسيم سكاني.

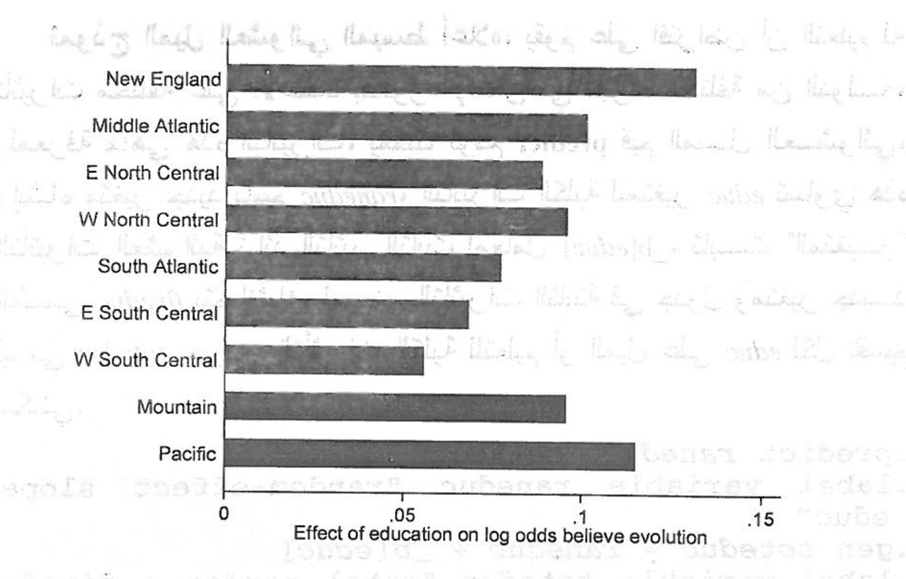
```
.predict raneduc, reffects
.label variable raneduc "random-effect slope
 educ"
.gen toteduc = raneduc + _b[educ]
.label variable toteduc "total random + fixed-
 effect slope educ"
.gen fixeduc = _b[educ]
```

```
.label variable fixeduc "fixed-effect slope
 educ (constant)"
.table cendiv, contents(mean fixeduc mean
 raneduc mean toteduc)
```

Census division	mean(fixeduc)	mean(raneduc)	mean(toteduc)
New England	.0924205	.0389457	.1313663
Middle Atlantic	.0924205	.0089432	.1013638
E North Central	.0924205	-.0036121	.0888085
W North Central	.0924205	.0035191	.0959396
South Atlantic	.0924205	-.0148976	.077523
E South Central	.0924205	-.0239878	.0684328
W South Central	.0924205	-.0366144	.0558061
Mountain	.0924205	.0033255	.095746
Pacific	.0924205	.0227141	.1151346

من الجدول يمكننا التأكيد بأن التأثير الكلي للتعليم يساوي التأثيرات الثابتة زائداً التأثيرات العشوائية، الشكل (10.13) يعرض هذه التأثيرات الكلية.

```
.graph hbar (mean) toteduc, over(cendiv)
 ytitle("Effect of education on log odds
 believe evolution")
```



الشكل (10.13)

نرى أن التعليم له تأثير موجب على الاحتمالات اللوغاريتمية لقبول فرضية تطور الإنسان في كل تقسيمات السكان، تأثيرات التعليم أكبر قوة بين المشاركين من ولايات نيوإنجلاند والمحيط الهادئ، مقارنة مع المناطق الوسطى الجنوبية الغربية والشرقية.



الفصل الرابع عشر

مقدمة في البرمجة

Introduction to Programming

كما رأينا سابقاً، فإنه يمكننا إنشاء نوع بسيط من البرامج بواسطة كتابة أي سلسلة من أوامر ستاتا في ملف نـصّي (ASCII)، أو بمحـرر الملـف التنفيذي Do-file Editor ببرنامج ستاتا (يمكنك الوصول إليه بالنقر على القائمة Window > Do-file Editor أو النقر على أيقونة) يعتبر طريقة سـهلة للقيام بذلك. بعد حفظ ملف do-file سوف نقوم بإدخال Stata، ونطبع أمرًا في شكل **do *filename*** فهذا يُخبر ستاتا بقراءة الملف المسمى *filename.do* وتنفيذ كل الأوامر التي يحتويها هذا الملف. كما يمكن إجراء عمليات برمجة أكثـر تعقيداً باستخدام لغة برمجة مدمجة ببرنامج ستاتا نفسه. أغلب أوامر سـتاتا التي تم استخدامها في الفصول السابقة، تتضمن برامج تم كتابتها في بـرامج ستاتا. هذه البرامج قد تم إنشاؤها أصـلاً بواسـطة شـركة سـتاتا أو مـن مستخدمين يحتاجون أشياء لا يمكن للغة برمجة ستاتا القيـام بهـا أو أنهـم يحتاجونها لمهام معينة.

برامج ستاتا يمكنها الوصول إلى كل مميزات سـتاتا، حيـث إن هـذه البرامج تتصل ببرامج أخرى، وهذه الأخيرة تتصل ببرامج أخرى، وتستخدم أدوات مساعدة لصياغة النماذج. وهذه الأدوات تتضمن جبر المـصفوفات، وتقدير الأرجحية العظمى، القدرة على كتابة برامج ستاتا توسّع مـا يمكننـا القيام به حتى ولوكان هدفنا واسعاً جداً مثل إضافة أساليب إحصائية جديدة أو هدف بسيط مثل إدارة قاعدة بيانات معينة.

البرمجة موضوع واسع في ستاتا، وهذا الفصل المختصر يعـرض مقدمـة لبعض المفاهيم والأدوات الرئيسة مع بعض الأمثلة عن كيفية استخدامها لتـسهيل

مهام تحليل البيانات. إذا كنت مهتماً بتعلم تفاصيل أكثر، فيمكنك الاطلاع على الدروس الموجودة بموقع ستاتا (www.stata.com/netcourse) فهي المكان المناسب كبداية، أما المرجع الرئيس حول البرمجة فهو دليل المستخدم للبرمجة *Programming Reference Manual* وجزءين من كتاب *Mata Matrix Programming Manual*، هناك تفاصيل عن تقدير الأرجحية العظمى والبرمجة في كتاب *Maximum Likelihood Estimation with Stata* (Gould, Pitblado and Poi 2010).

أدوات ومفاهيم أساسية : Basic Concepts and Tools

بعض الأدوات والمفاهيم الأساسية تُدمج مع قدرات برنامج ستاتا – تم شرحها في فصول سابقة – تعتبر كافية كبداية.

Do-files

الملفات التنفيذية do-files هي ملفات نصية من نوع (ASCII) تم إنشاؤها بواسطة محرر الملفات التنفيذية ببرنامج ستاتا Do-file Editor أو محرر نصوص أو أي برنامج تحرير نصي آخر. وهذه الملفات يتم حفظها بامتداد do. الملف يمكن أن يحتوي على أي سلسلة من أوامر ستاتا المنطقية. وفي ستاتا طباعة الأمر أدناه تقود برنامج ستاتا إلى قراءة الملف *filename.do* وتنفيذ الأوامر التي يحتويها هذا الملف:

```
.do filename
```

كل أمر في الملف *filename.do* بما فيه آخر أمر، يجب أن ينتهي بنهاية السطر ليبدأ من بداية سطر جديد مالم نضع محددًا من خلال الأمر **#delimit**

```
#delimit ;
```

هذا يضع فاصلة منقوطة كمحدد في نهاية السطر، وبذلك فإن برنامج ستاتا يعتبر أن السطر قد انتهى حتى يُصادف الفاصلة المنقوطة، ووضع الفاصلة المنقوطة كمحدد يسمح للأمر الواحد بأن يمتد لأكثر من سطر واحد، لاحقاً يمكننا ضغط مفتاح "إدخال" في لوحة المفاتيح كنهاية معتادة مع أمر **#delimit** آخر:

```
#delimit cr
```

ملاحظة مطبعية: العديد من الأوامر التي تظهر في هذا الفصل على الأرجح يتم طباعتها داخل ملف do-file بدلاً من طباعتها كأمر قائم بذاته في نافذة الأوامر، حيث تتم كتابة هذه الأوامر ضمن أوامر البرامج بدون عرض نقطة قبلها "." كما تم في المثالين السابقين أعلاه **delimit#** (ولكن ليس مع الأمر **do filename** والذي يجب كتابته في نافذة الأوامر كما هو معتاد).

Ado-files

ملفات Ado-files (التنفيذ الآلي) هي عبارة عن ملفات ASCII تحتوي على سلسلة من أوامر ستاتا مثل ملف do-files، الاختلاف بينها وبين ملفات do-file هو أننا لا نحتاج إلى طباعة الأمر ***filename*** **do** حتى نشغل الملف ado-file، بافتراض أننا قمنا بطباعة الأمر

```
.clear
```

كما هو الوضع مع أي أمر، فإن ستاتا يقوم بقراءة ذلك الأمر، وفحص ما إذا كان هناك أمر فعلي موجود بهذا الاسم. إذا كان الأمر **clear** غير موجود كجزء من أوامر ستاتا التنفيذية (وفي الحقيقة أنه موجود) فإن ستاتا سوف يبحث عن الأمر في قاموسه العادي وهو "ado" محاولاً إيجاد ملف باسم *clear.ado*، إذا وجد ستاتا الملف (كما يُفترض) فإنه يقوم بتنفيذ الأوامر التي يحتويها هذا الملف.

ملفات ado-files لها امتداد ado. والبرامج المكتوبة بواسطة المستخدمين (التي كتبتها أنت كمستخدم) في العادة يتم حفظها في مجلد باسم C:\ado\personal والبرامج المكتوبة بواسطة مستخدمي ستاتا الآخرين يتم حفظها في العادة في المجلد C:\ado\plus ومئات ملفات ado-files الرسمية يتم تثبيتها في المجلد C:\Program Files\Stata\ado، قم بطباعة **sysdir** لمشاهدة قائمة بالمجلدات المستخدمة من قبل برنامج ستاتا الحالي، وقم بطباعة الأمر **help sysdir** أو **help adopath** لمعرفة كيفية تعديلها.

الأمر **which** يوضح ما إذا كان أمرًا معيناً هو في الواقع أمر من أوامر ستاتا أو أمر موجود في ملف ado-file، وإذا كان الأمر هو أمر ado-file فيحدد مكانه، فمثلاً الأمر **summarize** من ضمن الأوامر المدمجة، ولكن

الأمر **regress** حالياً من ضمن الأوامر المعرّفة بملف ado-file والذي يُسمى regress.ado والذي تم تحديثه في أبريل 2011.

```
.which summarize
built-in command: summarize

.which regress
C:\Program Files\Stata\ado\base\r\regress.ado
 !*version 1.3.0  14apr2011
```

هذه التفرقة لا تمثل أي شيء لأغلب المستخدمين، لأن الأمر **summarize** والأمر **regress** يعملان بنفس السهولة عند استخدامهما. ودراسة الأمثلة واستعارة رمز من آلاف من ملفات ado-files ببرنامج ستاتا يمكن أن تساعدك عند البداية في كتابة برنامج ما، مخرجات الأمر **which** أعلاه تعطي موقع ملف *regress.ado* ولمشاهدة الرموز الفعلية في هذا الملف قم بطباعة الأمر

```
.viewsource regress.ado
```

ملفات ado-files تُعرّف أوامر التقدير ببرنامج ستاتا، وهذه الملفات تطورت بشكل ملحوظ، وأصبحت أكثر تعقيداً خلال السنوات الأخيرة، حيث إنها استوعبت قدرات جديدة ببرنامج ستاتا مثل **:svy**

البرامج : Programs

ملفات do-files وملفات ado-files قد يتم اعتبارها أنواعاً من البرامج. ولكن برنامج ستاتا يستخدم كلمة "برنامج" بمعناها الضيق لتعني سلسلة من الأوامر يتم حفظها في الذاكرة، وتنفيذها من خلال طباعة اسم برنامج معين، ملفات do-files أو ado-files أو الأوامر تطبع بشكل تفاعلي لتعريف مثل هذه البرامج، التعريف يبدأ مع عبارة تحدد اسم البرنامج، فمثلاً لإنشاء برنامج باسم *count5* نبدأ بطباعة

```
Program count5
```

الأسطر التالية يُفترض أن تحدد بشكل فعلي البرامج، وأخيراً نعطي أمر إنهاء **end** نتبعه بالضغط على مفتاح الإدخال.

```
end
```

عندما يقرأ ستاتا أوامر تعريف البرنامج، فإنه يحفظ تعريف البرنامج في الذاكرة، ويبدأ في تشغيله في أي وقت نطبع فيه اسم البرنامج كأمر:

```
.count5
```

البرامج تقوم بكفاءة بإنشاء أوامر جديدة متوافرة ببرنامج ستاتا، ولذا فإن أغلب المستخدمين لا يحتاجون إلى معرفة ما إذا كان أي أمر يأتي من برنامج ستاتا نفسه أو من برنامج ado-file.

ونحن في صدد البدء بكتابة برنامج جديد، فإننا عادةً نقوم بإنشاء إصدارات أولية ناقصة أو غير مكتملة، الأمر **program drop** مفيد ويسمح لنا بمسح برامج من الذاكرة حتى يمكننا تعريف إصدار جديد، فمثلاً لمسح برنامج *count5* من الذاكرة نقوم بطباعة الأمر

```
.program drop count5
```

لمسح كل البرامج (بدون مسح البيانات) من الذاكرة قم بطباعة الأمر:

```
.program drop _all
```

وحدات الماكرو المحلية : Local Macros

وحدات الماكرو عبارة عن أسماء (تصل إلى 31 حرفاً) يمكنها أن ترمز لسلاسل أو نتائج رقمية معرّفة ببرنامج أو قيم معرّفة للمستخدمين. وحدة الماكرو المحلية موجودة فقط مع البرامج التي تُعرّفها ولايمكن الإشارة إليها في برنامج آخر، ولإنشاء وحدة ماكرو محلية باسم *iterate* ترمز للرقم 0 قم بطباعة الأمر

```
    local iterate 0
```

وللإشارة إلى محتويات ماكرو محلي (0 في هذا المثال) قم بوضع اسم الماكرو بين علامة تنصيص فردية، فمثلاً

```
    display `iterate'
0
```

ولذا فإننا إذا كنا نريد زيادة قيمة *iterate* بقيمة واحد، فإننا نقوم بكتابة الأمر:

```
    local iterate = `iterate' + 1
    display `iterate'
1
```

بدلاً من رقم. فإن محتويات الماكرو يمكن أن تكون سلسلة نصية أو قائمة من الكلمات مثل:

```
local islands Iceland Faroes
```

ولمشاهدة محتويات سلسلة نصية يتم وضع علامات تنصيص مزدوجة حول اسم الماكرو الذي يجب أن يكون مُحاطاً بعلامة تنصيص مفردة:

```
    display "`islands'"
Iceland Faroes
```

يمكننا أن نضع سلسلة إضافية من الكلمات أو أرقام إلى محتويات الماكرو، فمثلاً

```
    local islands `islands' Newfoundland Nantucket
    display "`islands'"
Iceland Faroes Newfoundland Nantucket
```

قم بطباعة الأمر **help extended fcn** للحصول على معلومات أكثر عن دوال الماكرو الموسعة ببرنامج ستاتا حيث يقوم هذا الأمر بعرض معلومات عن محتويات وحدات الماكرو، فمثلاً يمكننا الحصول على عدد الكلمات في الماكرو، وحفظ هذا العدد كماكرو جديد باسم *howmany*:

```
    local howmany: word count `islands'
    display `howmany'
4
```

العديد من دوال الماكرو الموسعة الأخرى موجودة مع تطبيقات للبرمجة.

وحدات الماكرو الشاملة : Global macros

تشبه وحدات الماكرو الشاملة وحدات الماكرو المحلية، ولكن عند تحديدها فإنها تبقى في الذاكرة، ويمكن استخدامها بواسطة برامج أخرى خلال فترة استخدامك لبرنامج ستاتا. وللإشارة إلى محتويات الماكرو الشاملة سوف نبدأ باسم الماكرو مع علامة دولار (بدلاً من إرفاق الاسم في اليسار ويمين علامات الاقتباس كما تم مع وحدات الماكرو المحلية):

```
    global distance = 73
    display $distance * 2
146
```

ما لم نحدد بالضبط أننا نريد الاحتفاظ بمحتويات الماكرو لإعادة استخدامها لاحقاً، فإنه من الأفضل (أقل إرباكاً وأسرع في التنفيذ وأقل خطراً) استخدام ماكرو محلي بدلاً من ماكرو شامل في كتابة البرامج، ولحذف ماكرو من الذاكرة نقوم باستخدام الأمر macro drop.

```
macro drop distance
```

كما يمكننا حذف كل وحدات الماكرو من الذاكرة عن طريق الأمر:

```
macro drop _all
```

أوامر : Scalars

العدديات يمكن أن تكون أرقاماً أو سلاسل نصية يتم الإشارة إليها بواسطة اسم مثل وحدات الماكرو المحلية، ولاسترجاع محتوياتها لا نحتاج إلى إضافة اسم العدديّة ضمن علامات الاقتباس، فمثلاً:

```
scalar onethird = 1/3
display onethird
.33333333
display onethird*6
2
```

العدديات مفيدة جداً عند حفظ النتائج الرقمية للعمليات الحسابية بدقة كاملة، فالعديد من إجراءات برنامج ستاتا التحليلية تحتفظ بالنتائج مثل درجات الحرية، وإحصائيات الاختبار، والأرجحيات المسجلة وغيرها كعدديات يمكن مشاهدتها بطباعة الأمر return list أو ereturn list بعد إتمام عملية التحليل. العدديات ووحدات الماكرو المحلية والمصفوفات والدوال يتم حفظها تلقائياً بواسطة برامج ستاتا، وهي تمثل الأساسيات التي يمكن استخدامها في البرامج الجديدة.

الأمر : Version

قدرات برنامج ستاتا تغيرت خلال فترة من الزمن، وبالتالي فإن كتابة البرامج للإصدارات القديمة ببرنامج ستاتا قد لا تعمل بشكل مباشر مع الإصدار الحالي، الأمر version يعمل على حل هذه المشكلة حتى يمكن

استخدام البرامج القديمة، فعندما نحدد لبرنامج ستاتا الإصدار الذي تم استخدامه في كتابة البرنامج، فإن ستاتا يقوم بالتعديلات الضرورية حتى يمكن للبرنامج القديم العمل مع الإصدار الجديد لبرنامج ستاتا، فمثلاً إذا بدأنا البرنامج بالعبارة أدناه، فإن برنامج ستاتا يقوم باعتبار كل أوامر البرنامج كأنها مكتوبة بإصدار برنامج ستاتا رقم 9.

```
version 9
```

كتابة الأمر **version** في حد ذاته بدون أي إضافات تقوم بعرض الإصدار الحالي لبرنامج ستاتا.

التعليقات : Comments

لا يقوم برنامج ستاتا بمحاولة تنفيذ أي سطر يبدأ بعلامة نجمة، مثل هذه الأسطر يمكن استخدامها لإدراج تعليقات وشروحات في أي برنامج أو عرضها بشكل تفاعلي أثناء العمل على برنامج ستاتا، فمثلاً:

```
* This entire line is a comment.
```

وبدلاً من ذلك، يمكننا إدراج تعليق في السطر الذي يحتوي على الملف التنفيذي نفسه، وأبسط طريقة للقيام بذلك تتم بوضع التعليق بعد علامة // مزدوجة (مع مسافة واحدة على الأقل قبل علامة // المزدوجة)، فمثلاً

```
summarize logsize age // this part is the
     comment
```

كما يمكن استخدام علامة /// ثلاثية (يجب أن يسبقها مسافة واحدة على الأقل) تُشير إلى أن الذي يتبع هذه العلامات حتى نهاية السطر هو أمر، والسطر التالي هو عبارة عن أمر يجب تنفيذه كاستمرار للسطر الأول، فمثلاً:

```
summarize logsize age /// this part is the
        comment
educ income
```

حيث يتم تنفيذه وكأننا قمنا بطباعته كأمر:

```
summarize logsize age educ income
```

مع وجود تعليقات أو بدونها، فإن علامات /// الثلاثية تعني أن السطر التالي يجب قراءته كاستمرار للسطر السابق، فمثلاً السطرين أدناه سوف تتم قراءتهما كأمر **table** واحد حتى بعد فصلهما بواسطة الضغط على مفتاح الإدخال:

```
table married sex, ///
contents(median age)
```

علامات /// الثلاثية تُعتبر بديلاً لـ **#delimit;**، وهي طريقة تم شرحها سابقاً لكتابة أوامر البرامج التي تكون أطول من سطر واحد.

كما أنه من المحتمل إدراج تعليقات في منتصف سطر الأمر وذلك بوضعها بين علامة */ وعلامة /* فمثلاً:

```
table married sex, /* this is the comment */
  contents(median age)
```

إذا انتهى أحد الأسطر بعلامة */ فإن السطر التالي يبدأ بعلامة /* ثم يقوم برنامج ستاتا بتخطي الفاصل بين السطرين ويقرأ كلا السطرين وكأنهما أمر واحد، ويُفضل استخدام علامة ///

الحلقات : Looping

هناك عدد من طرق إنشاء حلقات البرامج، أحد أبسط هذه الطرق تستخدم الأمر **forvalues** فمثلاً البرنامج أدناه يقوم بالعد من 1 إلى 5.

```
* Program that counts from one to five
program count5
   version 12.1
   forvalues i = 1/5 {
      display `i'
   }
end
```

بطباعة هذه الأوامر نحن نقوم بتعريف البرنامج *count5* وبدلاً من ذلك يمكننا استخدام محرر الملف التنفيذي Do-File Editor لحفظ سلسلة من الأوامر على شكل ملف ASCII باسم *count5.do* ثم نقوم بطباعة الأمر أدناه الذي يجعل برنامج ستاتا يقرأ الملف

```
.do count5
```

الطريقة الأخرى تتم من خلال تعريف البرنامج *count5* وذلك بجعل هذا المتغير كأنه أمر جديد

```
.count5
1
2
3
4
5
```

الأمر:

```
forvalues i = 1/5 {
```

يقوم بتخصيص وحدة ماكرو محلية *i* لأعداد صحيحة متتالية من 1 وحتى 5، الأمر هو

```
display `i'
```

الأمر أعلاه يعرض محتويات الماكرو، اسم الماكرو *i* تم اختياره عشوائياً، وهناك تركيبة أخرى للأمر تسمح لنا بالعد من 0 إلى 100 بفرق 5 (0، 5، 10، ... 100):

```
forvalues j = 0(5)100 {
```

الخطوات بين القيم ليس بالضرورة أن تكون أعدادًا صحيحة طالما النهاية واحدة، وللقيام بالعد من 4 إلى 5 باستخدام زيادة 0.01 (4.00، 4.01، 4.02 ... 5.00) نقوم بكتابة:

```
forvalues k = 4(.01)5 {
```

أي سطر يحتوي على أوامر ستاتا صحيحة بين أقواس البداية والنهاية { } يتم تنفيذها بطريقة متكررة لكل القيم المحددة، وبغض النظر عن التعليقات الاختيارية فلا شيء في سطر الأمر يتبع قوس البداية، أما قوس النهاية فيتطلب أن يكون هناك سطر خاص به.

الأمر **foreach** يستخدم طريقة مختلفة، فبدلاً من تحديد مجموعة من القيم الرقمية التنفيذية، فإننا نقوم بإعطاء قائمة بالعناصر التي يحدث بها التكرار،

هذه العناصر يمكن أن تكون متغيرات أو ملفات أو سلاسل نصية أو قيم رقمية. ولمعرفة كيفية تركيبة هذا الأمر قم بطباعة **help foreach**.

الأمر **forvalues** والأمر **foreach** تقوم بإنشاء حلقات وتقوم هذه الحلقات بتكرار أرقام محددة مسبقاً لعدة مرات، إذا كنا نريد الحلقات أن تستمر حتى تحقق شروطاً معينة، فإن الأمر **while** سوف يكون مفيدًا في ذلك. جزء من البرنامج مع الشكل العام أدناه سوف يقوم بشكل متكرر بتنفيذ الأوامر الموجودة بين الأقواس المعكوفة { } هذا التكرار سوف يستمر طالما أن تقييم برنامج *expression* "صحيح" "true".

```
while expression {
    command A
    command B
    . . . .
}
command Z
```

كما رأينا في المثال السابق، فإن قوس الإقفال } يجب أن يكون في سطر منفصل خاص به وليس في نهاية سطر أي أمر.

عندما يكون تقييم *expression* "خطأ" "false" فإن الحلقات تتوقف ويقوم برنامج ستاتا بتنفيذ الأمر *Z*، وبنفس ما قمنا به في المثال السابق، فإن التالي برنامج مبسط يقوم باستخدام حلقة **while** التي تعرض على الشاشة تكرار الأرقام من 1 إلى 6:

```
* Program that counts from one to six
program count6
    version 12.1
    local iterate = 1
    while `iterate' <= 6 {
        display `iterate'
        local iterate = `iterate' + 1
    }
end
```

الحلقة الأكثر أهمية تظهر في برنامج *multicat.ado* سوف يـتم شـرحه لاحقاً في هذا الفصل. وللحصول على معلومات أكثر عن ذلك قم بـالاطلاع على دليل المستخدم *Programming Referen :e Manual*.

If … else

الأمر **if** والأمر **else** يحددان لبرنامج ما أن يقوم بشيء واحد إذا كـان الشرط *expression* صحيحاً ويقوم بشيء آخر إذا لم يتوافر هذا الشرط، وتتم كتابة هذا الشرط كما يلي:

```
if expression {
    command A
    command B
    . . . .
}
else {
    command Z
}
```

فعلى سبيل المثال، هناك جزء في البرنامج أدناه يقوم بفحـص مـا إذا كانت محتويات الماكرو المحلي *span* عـددًا فرديـاً وإبـلاغ المـستخدمين بالنتيجة.

```
if int(`span'/2) != (`span' - 1)/2 {
   display "span is NOT an odd number"
}
else {
   display "span IS an odd number"
}
```

الشروط : Arguments

البرامج تحدد الأوامر الجديدة، في بعض الأمثلة (كما كان الوضع فـي المثال السابق **count5**) قررنا أن الأمر سوف يقوم بنفس الشيء بالضبط في كل مرة يتم استخدامه، وفي الغالب نحن نحتاج إلى أمر يتم تغييره بواسـطة شروط مثل أسماء المتغيرات أو الخيارات. هناك طريقتان يمكننـا بهمـا أن نحدد لبرنامج ستاتا كيف يقرأ ويفهم سطر الأمر الذي يحتوي على شـروط، أسهل هذه الطرق هي استخدام الأمر **args**.

الملف التنفيذي do-file أدناه (*listres1.do*) يُعرّف برنامج يقوم بحساب انحدار متغيرين، ثم يضع المشاهدات في قائمة مع أكبر بواقي مطلقة، الأمر **listres1** يعرض عدة أشياء خاطئة مثل استبعاد متغيرات وترك متغيرات جديدة أخرى في الذاكرة، والتي يمكن أن يكون لها تأثيرات جانبية غير مرغوبة. وعموماً فإن هذا الأمر يساعد في توضيح استخدام المتغيرات المؤقتة.

```
* Perform simple regression and list
  observations with #
     * largest absolute residuals.
* syntax: listres1 Yvariable Xvariable #
  IDvariable
     capture drop program listres1
     program listres1, sortpreserve
       version 12.1
       args Yvar Xvar number id
       quietly regress `Yvar' `Xvar'
       capture drop Yhat_
       capture drop Resid_
       capture drop Absres_
       quietly predict Yhat_
       quietly predict Resid_, resid
       quietly gen Absres_ = abs(Resid_)
       gsort -Absres_
       drop Absres_
       list  `id'  `Yvar'  Yhat_  Resid_  in
       1/`number'
end
```

السطر **args Yvar Xvar number id** يحدد لبرنامج ستاتا شروطاً لأربع وحدات ماكرو. هذه الشروط يمكن أن تكون أرقاماً أو أسماء متغيرات أو سلاسل نصية أخرى يتم الفصل بينها بمسافة. أول شرط هو استخدام محتويات ماكرو محلي اسمه *Yvar* والثاني ماكرو محلي اسمه *Xvar* وهكذا، بعد ذلك يقوم الأمر باستخدام محتويات وحدات الماكرو هذه في أوامر أخرى مثل الانحدار:

```
quietly regress `Yvar' `Xvar'
```

البرنامج يقوم بحساب قيم البواقي المطلقة (*Absres*) ثم يستخدم الأمر **gsort** (مع علامة ناقص قبل اسم المتغير) لترتيب البيانات من أعلى إلى أسفل مع وضع القيم المفقودة في الأخير:

```
gsort -Absres_
```

الخيار **sortpreserve** في سطر الأمر يجعل هذا البرنامج "يحافظ على الترتيب" مؤكداً على أن ترتيب المشاهدات هو نفسه بعد تشغيل البرنامج كما كان قبل التشغيل.

ملف البيانات *Nations2.dta* يحتوي على بيانات 194 دولة تتعلق بنسبة انبعاث ثاني أكسيد الكربون CO_2 لكل شخص (*co2*) والناتج المحلي الإجمالي لكل شخص (*gdp*) واسم الدولة (*country*)، يمكننا فتح هذا الملف واستخدامه لتوضيح البرنامج الجديد، الأمر **do** يقوم بتشغيل الملف التنفيذي do-file المُسمى *listres1.do* وبهذه الطريقة يُعرّف البرنامج الأمر الجديد **listres1**:

```
.use C:\data\Nations2.dta, clear
.do C:\data\listres1
```

ثم بعد ذلك نقوم باستخدام الأمر الجديد الذي تم تعريفه وهو **listres1** يتبع شروطه الأربعة. الشرط الأول: يحدد المتغير *y* والثاني: *x* والثالث: كم عدد المشاهدات التي يجب وضعها في قائمة، والرابع: يُعطي هذه الحالة رقماً خاصاً؛ في هذا المثال أدناه، الأمر يقوم بإعداد قائمة بالمشاهدات التي لها أكبر قيم بواقي مطلقة.

```
.listres1 co2 gdp 5 country
```

	country	co2	Yhat_	Resid_
1.	Qatar	210.65	114.4057	96.2443
2.	Bahrain	102.65	45.54433	57.10566
3.	Trinidad/Tobago	89.25	34.18739	55.06261
4.	Kuwait	118.2	67.83949	50.36051
5.	United Arab Emirates	120.85	79.20002	41.64998

في هذه الدول الخمس المصدّرة للنفط انبعاث ثاني أكسيد الكربون لكل فرد أعلى من المتوقع.

الأمر : Syntax

الأمر syntax يُعتبر طريقة معقدة، ولكن أيضاً مفيدة لقراءة سطر أي أمر، الأمر التنفيذي أدناه والمسمى *listres2.do* يُشبه الأمر السابق، ولكن يستخدم الأمر syntax بدلاً من الأمر args.

```
*Perform simple or multiple regression and list
*observations with # largest absolute
 residuals.
*listres2 yvar xvarlist [if] [in], number(#)
 [id(varname)]
capture drop program listres2
program listres2, sortpreserve
version 12.1
syntax varlist(min=1) [if] [in],
Number(integer) [Id(varlist)]
   marksample touse
   quietly regress `varlist' if `touse'
   capture drop Yhat_
   capture drop Resid_
   capture drop Absres_
   quietly predict Yhat_ if `touse'
   quietly predict Resid_ if `touse', resid
   quietly gen Absres_ = abs(Resid_)
   gsort -Absres_
   drop Absres_
   list `id' `1' Yhat_ Resid_ in 1/`number'
end
```

الأمر listres2 له نفس وظيفة الأمر listres1 حيث يقوم بحساب الانحدار، ثم يقوم بإنشاء قائمة للمشاهدات مع أكبر بواقي مطلقة، هذا الإصدار الجديد من الأمر، يحتوي على عدة تحسينات، والتي أمكن الحصول عليها عن طريق الأمر syntax، والأمر الجديد غير مقيّد بانحدار ذي متغيرين كما حدث في الأمر listres1. الأمر listres2 سوف يعمل مع أي عدد من المتغيرات التنبؤية بما فيها تلك التي تكون قيمها المتوقعة تساوي متوسط متغيرات y وبواقيها هي انحرافات من المتوسط. الأمر listres2 يسمح بخيارات if وin، كما أن استخدام متغير ما يُحدّد المشاهدات، وهو اختياري مع الأمر listers2 بدلاً من أن تكون المشاهدات مطلوبة كما كان

الوضع مع الأمر listres1، فمثلاً يمكننا حساب انحدار انبعاث ثاني أكسيد الكربون على الإنتاج المحلي الإجمالي، ونسبة المناطق الحضارية، مع حصر التحليل ليشمل الدول في المنطقة 2 فقط وهي دول أمريكا.

```
.do C:\data\listres2.do
.listres2 co2 gdp urban if region == 2, n(5)
 i(country)
```

	country	co2	Yhat_	Resid_
1.	Trinidad/Tobago	89.25	47.63852	41.61148
2.	Barbados	16.65	35.0574	-18.40739
3.	Saint Kitts/Nevis	10.05	26.28106	-16.23106
4.	Antigua and Barbuda	18.3	34.44279	-16.1428
5.	Suriname	19.45	5.137903	14.3121

سطر الأمر **syntax** في هذا المثال يوضح بعض السمات العامة للأمر:

```
syntax varlist(min=1) [if] [in], Number
(integer) [Id(varlist)]
```

قائمة المتغيرات للأمر **listres2** يجب أن تحتوي على اسم متغير واحد على الأقل (**(varlist(min=1)**)، الأقواس تشير إلى شروط اختيارية، وهي في هذا المثال المحددات **if** و**in** والخيار **id( )**، والحروف الكبيرة الأولى للخيارات تشير إلى أقل مشاهدة يمكن استخدامها، وحيث إن سطر **syntax** في هذا المثال تم تحديده **Number(integer) Id(varlist)** والأمر الواقعي يمكن كتابته كما يلي:

```
.listres2 co2 gdp, number(6) id(country)
```

أو يمكن كتابته كما يلي

```
.listres2 co2 gdp, n(6) i(country)
```

محتويات الماكرو المحلي *number* يجب أن تكون عددًا صحيحاً و*id* هي أسماء متغير واحد أو أكثر.

هذا المثال يشرح أيضاً الأمر **marksample** والذي يجعل العينة الفرعية (مثل التي تم تحديدها بواسطة **if** و**in**) قابلة للاستخدام في تحليلات لاحقة.

تركيبة الأمر syntax تم تلخيصها فــي دليــل المــستخدم *Programming Manual*، فتجربة ودراسة البرامج الأخرى تساعد في فهم هذا الأمر.

أمثلة عن البرامج - برنامج multicat (الرسم البياني للعديد من المتغيرات النوعية) : Example Program: multicat (Plot Many Categorical Variables)

في الأجزاء السابقة، تم شرح أفكار بسيطة وأمثلة عن برامج قــصيرة. في هذا الجزء سوف نطبق هذه الأفكار على بــرامج أكبــر تحــدد إجــراء إحصائياً جديدًا باسم **multicat**.

وبيانات وإجراءات البحث الاستقصائي تتضمن العديد مــن المتغيــرات النوعية أحياناً قد تصل إلى 100 متغير أو أكثر، مقتطفات من الاســتطلاع الاجتماعي العام لسنة 2010 توفر لنا مثالاً بسيطاً مع 19 متغيــرًا أغلبهــا ردود نوعية على أسئلة الاستطلاع.

```
.use C:\data\GSS_2010_SwS.dta, clear
.describe

Contains data from C:\data\GSS_2010_SwS.dta
  obs:           809                     General Social Survey 2010--evolution
                                           etc.
 vars:            23                     6 Mar 2014 01:41
 size:        34,787
------------------------------------------------------------------------------
              storage  display    value
variable name   type   format     label      variable label
------------------------------------------------------------------------------
id              int    %8.0g                 Respondent ID number
year            int    %8.0g                 GSS year
wtssall         float  %9.0g      LABCM      probability weight
cendiv          byte   %15.0g     cendiv     Census division
logsize         float  %9.0g                 log10(size place in 1,000s, +1)
age             byte   %8.0g      age        Age in years
nonwhite        byte   %9.0g      nonwhite   Consider self white/nonwhite
sex             byte   %8.0g      sex        Respondent gender
educ            byte   %8.0g      educ       Highest year of schooling
married         byte   %9.0g      yesno      Currently married
income06        byte   %15.0g     income     Total family income
polviews        byte   %12.0g     polviews   Polit views liberal-conservative
bush            byte   %9.0g      yesno      Voted for Bush in 2004
obama           byte   %9.0g      yesno      Voted for Obama in 2004
postlife        byte   %8.0g      yesno      Believe in life after death
grass           byte   %9.0g      grass      Should marijuana be legalized?
gunlaw          byte   %9.0g      gunlaw     Oppose permit required to buy gun
sealevel        byte   %10.0g     sealevel   Bothered if sea level rose 20 ft
evolve          byte   %9.0g      true       Humans developed/ earlier species
recendiv        float  %9.0g                 random-effect intercept cendiv
raneduc         float  %9.0g                 random-effect slope educ
toteduc         float  %9.0g                 total random + fixed-effect slope educ
fixeduc         float  %9.0g                 fixed-effect slope educ (constant)
------------------------------------------------------------------------------
Sorted by:  id
```

كخطوة أولى في استكشاف مثل هذا النوع من البيانات أو إعداد تقرير تمهيدي، فإننا قد نقوم بإعداد جداول تعرض نسبة التوزيعات لكل متغير، الأمر أدناه سوف يقوم بإنشاء ثمانية جداول لكل المتغيرات الموجودة في ملف البيانات من المتغير *polviews* وحتى المتغير *evolve*.

```
.tab1 polviews - evolve
```

عموماً برنامج ستاتا لايوفر طريقة سهلة لإنشاء وحفظ الرسم البياني للأعمدة لقائمة متغيرات، وكمثال عن البرمجة، فإن هذا الجزء يعرض برنامجاً مؤقتاً تمت كتابته لمقابلة احتياج معين عند العمل مع بحوث الدراسات الاستطلاعية الأكثر تعقيداً.

برنامج **multicat** – والذي تم تعريفه بواسطة ملف do-file أدناه – والذي تم إنشاؤه بناءً على برنامج اسمه **catplot** كتبه مستخدم آخر تم شرحه في الفصل 4، الأمر **catplot** يمكنه رسم عدد متنوع من الأشكال البيانية التي تعرض توزيعاً متغيرًا نوعياً، برنامج **multicate** متخصص أكثر في إنشاء الرسم البياني للأعمدة الأفقية، بحيث تكون هناك نسبة مئوية لكل فئة. ولكن هذا التنسيق مفيد لعرض بيانات الدراسات الاستقصائية، برنامج **multicat** له القدرة على العمل مع قائمة بها عدة متغيرات لا يمكن لـ **catplot** وأوامر الرسم البياني الأخرى ببرنامج ستاتا التعامل معها. وبالتالي، يمكننا أن نستخدم برنامج **multicat** في رسم أعمدة بيانية أفقية لكل متغير في البيانات الموجودة لدينا، وحفظ كل شكل بياني على حدة. البرنامج يحفظ الأشكال البيانية بتنسيقين اثنين، الأول بتنسيق ستاتا الذي له امتداد (gph)، والثاني بواحد من عدة تنسيقات (emf, eps, pdf بناءً على نظام التشغيل) وأسماء هذه الملفات سوف يكون بناءً على أسماء المتغيرات، ويمكنك تغيير أي من هذه المواصفات، وذلك من خلال تحرير الملف *multicat.ado*، وتعديل البرنامج بطريقة تلائم احتياجاتك التحليلية.

```
*! version 2.0 21jun2012
*! L. Hamilton, Statistics with Stata (2012)
* Requires catplot.ado installed. Graphs are
  saved in default directory.
program define multicat
```

```
version 12.1
   syntax varlist [if] [in] [aweight fweight
   iweight] ///
   [, MISSing BY(varname) OVER(varname) ]
   if "`over'" != "" {
      display as error "over() option not
      allowed with multicat;"
      display as error "use by() option or try
      catplot command
      instead."
      exit 198
   }
   marksample touse, strok novarlist
   if "`weight'" != "" local Weighted_ =
   "Weighted"
   if "`c(os)'"=="Windows" {
      local filetype "emf"
   }
   else if "`c(os)'"=="Unix" {
      local filetype "eps"
   }
   else if "`c(os)'"=="MacOSX" {
      local filetype = "pdf"
   }
   else {
      display as error "unknown operating
      system: `c(os)'"
      exit 799
   }
   capture {
      if "`by'" != "" {
         foreach var of varlist `varlist' {
            local Vlab_: variable label `var'
            catplot hbar `var' [`weight'
            `exp'] if `touse', ///
            blabel(bar, format(%3.0f)) ///
            percent(`by') ytitle("`Weighted_'
            Percent") ///
            `missing' by(`by',
            title("`Vlab_'", size(medium)))
         graph save -`by'-`var'.gph, replace
         graph export -`by'-`var'.`filetype',
         replace
      }
   }
   else {
```

```
        foreach var of varlist `varlist' {
         quietly tab `var' if `touse', `missing'
         local Nofobs_ = r(N)
         local Vlab_: variable label `var'
         catplot hbar `var' [`weight' `exp'] if
         `touse', ///
           blabel(bar, format(%3.0f)) ///
           percent  ytitle("`Weighted_'  Percent,
           N = `Nofobs_'") ///
           title("`Vlab_'",        size(medium))
           `missing' `options'
         graph save Graph -`var'.gph, replace
         graph export -`var'.`filetype', replace
       }
     }
   }
   error _rc
end
```

المسافة في بداية الأسطر ليس لها تأثير على تنفيذ البرنامج، ولكن تجعل قراءة البرامج أسهل للمبرمجين، الشيء الجوهري في برنامج **multicat** هو تركيبة جملته **syntax**، ومن ثم حلقة **foreach** التي تقوم بشكل متكرر بإعطاء أمر **catplot** لكل متغير في قائمة المتغيرات. وحدات الماكرو المحلية ترسل معلومات إلى الأمر **catplot** الذي يقوم بدوره برسم الأشكال البيانية، الأمر يسمح بأوزان تحليلية والتي لها هنا تأثير يشبه تأثير الأوزان الاحتمالية في الأمر **svy:tab**، كما يسمح كذلك بالمحددات **if** و**in**، وبشكل اختياري يمكننا إدراج القيم المفقودة **missing** واستخدام **by()** ولكن لايمكننا استخدام **over()**.

برنامج **multicat** تمت كتابته تدريجياً، حيث تم البدء مع ملف do-file اسمه *multicat.do* وكانت البداية مع إدخال المكونات مثل جملة التركيب، ثم بعد ذلك تشغيل هذا الملف لمشاهدة كيف يعمل قبل إضافة أي مكونات أخرى، ويجب ملاحظة أن تشغيل الملفات التجريبية لا يقوم بإخراج نتائج مرضية، قم بطباعة الأمر

```
.set trace on
```

حيث يأمر برنامج ستاتا بعرض البرامج خط بخط بنفس ترتيب تنفيذها، وبذلك يمكننا مشاهدة أين يحدث الخطأ بالضبط، ولاحقاً يمكننا أن نقوم بإيقاف هذه الميزة بطباعة الأمر

```
.set trace off
```

النسخة التمهيدية لملف *multicat.do*، فإن السطر الأول يحتوي على **capture program drop *multicat*** مهمته حذف البرنامج من الذاكرة قبل تعريفه مرة أخرى، وهذا ضروري في مرحلة الكتابة والتصحيح أوعندما تكون النسخة السابقة من البرنامج ناقصة أو غير صحيحة، وعموماً فإن مثل هذا السطر يجب حذفه عند إتمام كتابة البرنامج.

عندما نعتقد بأن ملف do-file يقوم بتعريف برنامج نرغب في استخدامه مرة أخرى، فيمكننا إنشاء ملف ado-file، ويمكن القيام بذلك من خلال حفظ الملف مع امتداد (*multicat.ado*)ado.، ويُفضّل حفظ الملف في المجلد ado\personal وقد تحتاج إلى إنشاء هذا المجلد إذا لم يكن موجودًا مسبقاً، يمكن الحفظ في المجلدات الأخرى، ولكن يجب مراجعة دليل المستخدم *User's Manual* لمعرفة أين يبحث ستاتا عن ملفات ado-files؟ قبل الاستمرار، وعند إتمام ذلك فيمكننا استخدام **multicat** كأمر اعتيادي ضمن برنامج ستاتا.

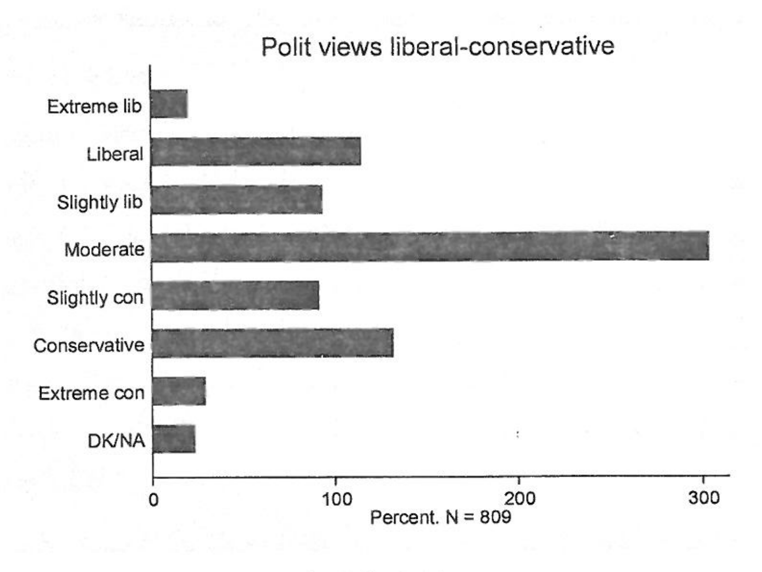
يمكن تحسين البرنامج لجعله أكثر مرونة وأناقة وسهولة، ويجب ملاحظة أن تضمين الأوامر مصدر البرنامج "version 2.0" في أول سطرين واللذين يبدآن بعلامتي !* فهذا الأمر يشير إلى الإصدار الثاني من ملف *multicat.ado* وليس إصدار ستاتا (الإصدار السابق من ملف *multicat.ado* يظهر في الإصدار السابق من هذا الكتاب)، إصدار ستاتا المناسب لتشغيل هذا البرنامج تم تحديده على أساس أنه الإصدار **version 12.1** والذي يظهر في سطور لاحقة في البرنامج، بالرغم من أن الأوامر التي تبدأ !* لا تؤثر على الطريقة التي يعمل بها البرنامج، إلا أنه يمكن مشاهدتها بواسطة الأمر **which**

```
. which multicat
.\multicat.ado
*! version 2.0 21jun2012
*! L. Hamilton, Statistics with Stata (2012)
```

استخدام برنامج Multicat : Using Multicat

بعـــد حفـــظ الملـــف *multicat.ado* (مـــثلاً تـــم حفظـــه فـــي المجلـــد C:\ado\personal) فإن الأمر **multicat** يصبح قابلاً للاستخدام كأنه أمر مـــن أوامر ستاتا العادية (حتى وإن لم تكن مكتملـــة)، الـــشكل (1.14) يعـــرض إجابات بخصوص وجهات نظر سياسية، النسب والأرقام للمشاهدات تظهـــر في الرسم البياني.

```
.multicat polviews, missing
```



الشكل (1.14)

بيانات الاستطلاع يتم تحليلها عمومـــاً باســـتخدام الأوزان الاحتماليـــة، ويجب أن يتم تحديد أن البيانات هي بيانات استقصائية باستخدام svyset كمـــا سبق، وإن تم شرحه فـــي الفـــصل (4)، تطبيـــق الأوزان علـــى البيانـــات الاستقصائية svy: tab يوضح أن الردود أدناه للإجابات عن تشريع اســـتخدام مخدّر المارجوانا.

```
.svy: tab grass, percent miss
```

```
(running tabulate on estimation sample)

Number of strata   =        1              Number of obs     =        809
Number of PSUs     =      809              Population size   = 812.73293
                                           Design df         =        808
```

Should marijuana be legalized ?	percentages
Not	29.68
Legal	30.04
DK	5.035
NA	35.24
Total	100

```
Key:  percentages  =  cell percentages
```

هذا السؤال له نوعان من القيم المفقودة، حوالي 5% من المشاركين في الدراسة تم سؤالهم عن مخدّر الماريجوانا، ولكنهم أجابوا بأنهم لا يعرفونه، هذه القيم المفقودة تم ترميزها بـ a. مع توصيفها بـ"DK"، وحوالي نسبة 35% في هذه العينة لم يتم سؤالهم عن رأيهم في تشريع استخدام الماريجوانا وهؤلاء تم ترميزهم بالرمز b. مع توصيف للقيم بـ "NA"؛ بيانات *grass* GSS تحتوي على عدد كبير من الأسئلة، حيث بها أسئلة مختلفة لعينات فرعية، ومن ناحية تحليلية، فإن ذلك يجعل للبيانات معنى، حيث يمكننا استبعاد المجموعة التي لم يتم سؤالها أسئلة معينة وحساب النسب، وبذلك نجد أن هناك انقساماً متساوياً: نحو 46% لصالح تشريع الماريجوانا، و46% ضد مع وجود نسبة حوالي 8% لم يتخذوا قرارًا.

```
.svy: tab grass if grass< .b, percent miss
```

```
(running tabulate on estimation sample)

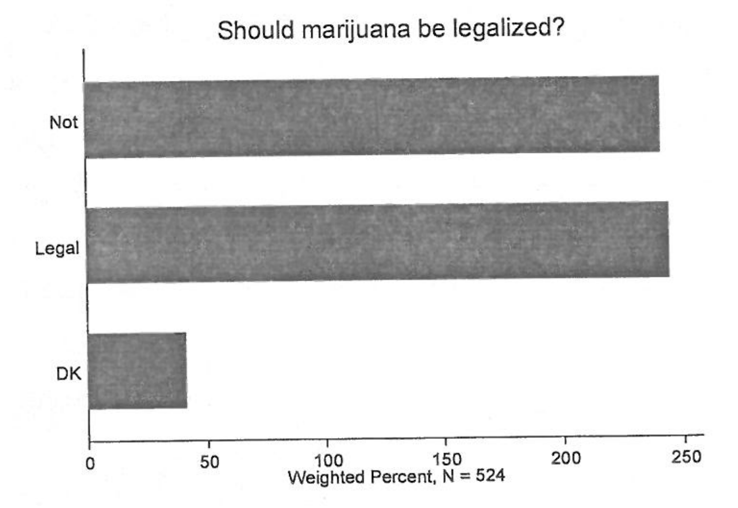
Number of strata   =          1                Number of obs      =        524
Number of PSUs     =        524                Population size    = 526.30952
                                               Design df          =        523
```

Should marijuana be legalized ?	percentages
Not	45.83
Legal	46.39
DK	7.776
Total	100

```
Key:  percentages  =  cell percentages
```

الأمر **multicat** (الذي تم إنشاؤها بناءً على الأمر **catplot**) لا يفهم أوامر **svy:** أو الأوزان الاحتمالية، ولكن الأوزان التحليلية لها نفس التـأثير هنـا، الشكل (2.14) يعرض شكلاً بيانياً للأمر **multicat** يتعلق بالجـدول أعـلاه، ويُلاحظ بأن الشكل البياني للأمر **multicat** لاحظ بأن حجم العينة (524).

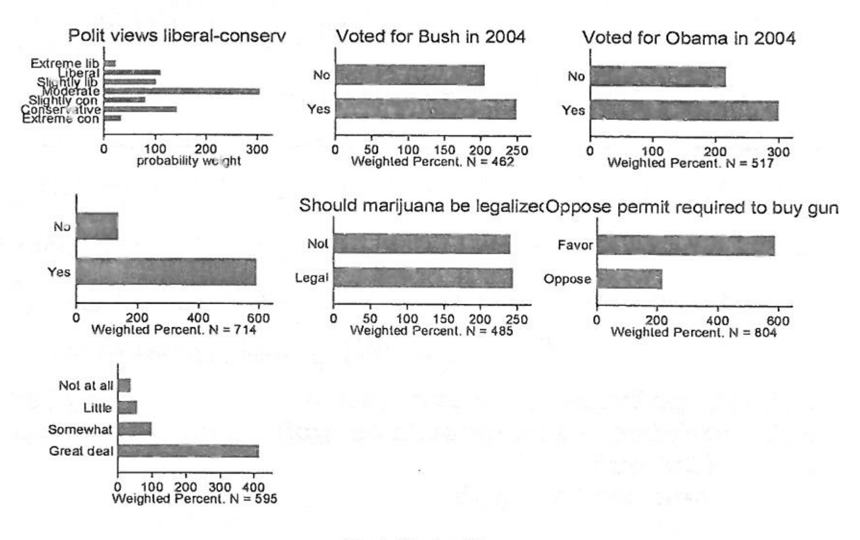
```
.multicat grass if grass< .b [aw = wtssall], miss
```



الشكل (2.14)

هناك سلسلة من عدة أشكال بيانية تشبه الشكل (2.14) تم استبعادها في المستند أو الشرائح المعروضة ويمكن قراءتها وإضافة ملاحظات عليها بواسطة المحلل، وذلك لإجراء عرض سريع للنتائج. وبغض النظر عن التعقيد المصاحب للقيم المفقودة لدينا، هنا مثال سريع يمكننا من خلاله استخدام الأمر **multicat** لرسم 8 أعمدة بيانية للمتغيرات من المتغير *polviews* إلى المتغير *evolve*. حيث يقوم الأمر بحفظ كل شكل بياني بطريقة آلية مع أسماء ملفات مثل *polviews.gph*، ثم يتم بعد ذلك استخدام الأمر **graph combine** لدمج الأشكال البيانية معاً في صورة واحدة لتكوّن الشكل (3.14).

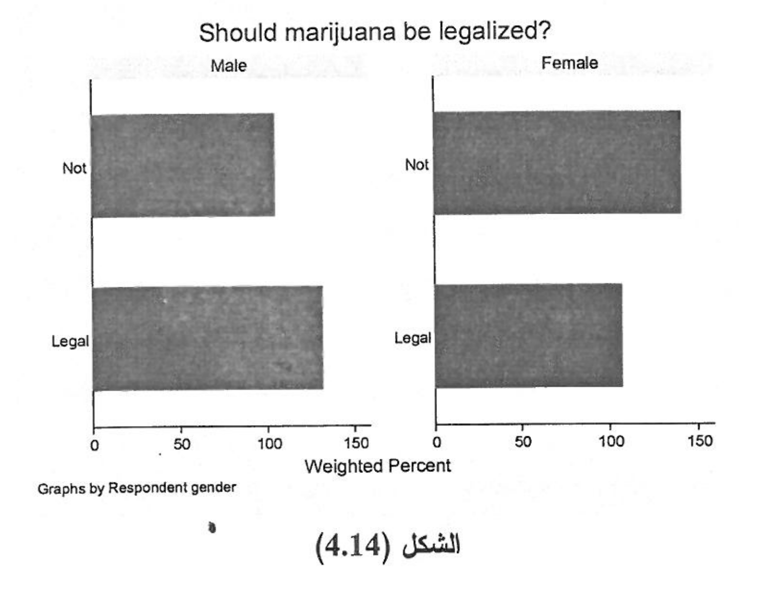
```
.multicat polviews-evolve [aw = wtssall]
.graph combine -polviews.gph -bush.gph -
 obama.gph -postlife.gph
    -grass.gph -gunlaw.gph -sealevel.gph -
    evolve.gph
```



الشكل (3.14)

البحوث الاستقصائية تصبح أكثر إثارة عندما نقارن بين المجموعات الفرعية، فمثلاً الفصل المنتظم حول تشريع المارجوانا يظهر بشكل مختلف عندما نقوم بتقسيم الإجابات على أساس الجنس في الشكل (4.14).

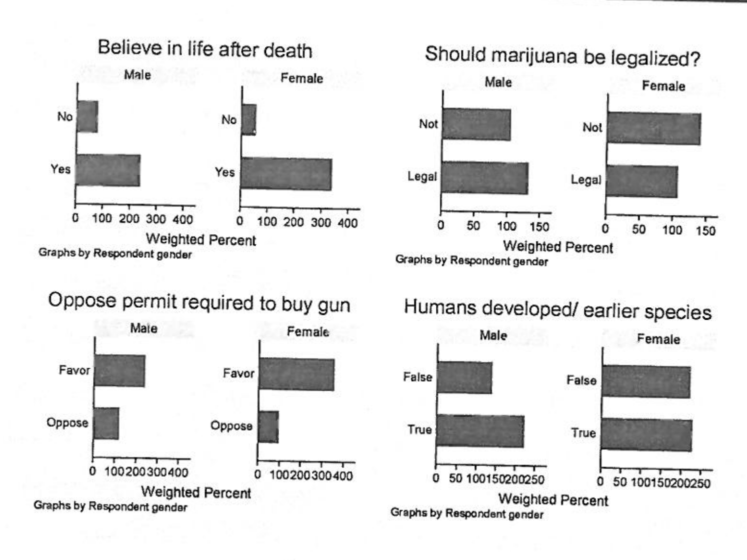
```
.multicat grass [aw = wtssall], by(sex)
```



الشكل (4.14)

وفي هذا الشكل، يعرض الأمر **multicat** توضيحاً أكثر عند المقارنة بين الأشكال البيانية، الأوامر أدناه تقوم بإنشاء 8 أشكال بيانية للآراء وتقسيم هذه الآراء بناءً على الجنس ثم تجميع 4 أشكال بيانية في شكل واحد تظهر في الشكل (5.14)، نرى أن ردود الإناث أكثر اعتقاداً بأن هناك حياة بعد الموت (86 مقابل 76%) كما أنهن يعارضن تشريع تعاطي المارجوانا (57 مقابل 43%) وهن يساندن تصريح استخدام السلاح (77 مقابل 68%) ويرفضن الآراء المتعلقة بالتطور البشري (49 مقابل 39%).

```
.multicat polviews-evolve [aw = wtssall], by(sex)
.graph combine -sex-postlife.gph -sex-grass.gph -
 sex-gunlaw.gph
        -sex-evolve.gph
```



الشكل (5.14)

تم استخدام الأمر **graph combine** في الشكلين (3.14) و(5.14) لـدمج الأشكال البيانية وتوضيح ماقام به الأمر **multicat**. وفي البحوث، فإن العدد الفعلي للأشكال البيانية عادة يفوق ما نريد إدراجه في شكل بيـاني واحـد، والأمر **multicat** يمكنه بكل سهولة رسم 100 شكل بياني، ومقارنة إجابـات المشاركين في الدراسات الاستقصائية من حيث الجنس ثم إجراء 100 مقارنة أخرى من حيث مستوى التعليم والفئة العمرية والانتماء السياسي والجغرافي وأي تصنيفات أخرى لها أهمية للباحث، أغلب عمليات التحليل سـوف لـن تحتاج إلى هذا الأمر الخاص، ولكن عندما تكون هناك احتياجات معينة فـي بحث ما، فإن البرامج المؤقتة من هذا النوع قد تكون ضرورية.

ملف المساعدة : Help File

ملفات المساعدة هي سمة أساسية عند استخدام برنامج ستاتا، فـالبرامج المكتوبة بواسطة المستخدمين مثل *multicat.ado* يمكن أن تصبح أكثر أهمية لعدم وجود توثيق خاص بها في دليل المستخدم، ويمكننا كتابة ملف مساعدة لبرنامج *multicat.ado* بواسطة استخدام Do-file Editor لإنشاء ملـف نـصّي

باسم *multicat.sthlp* يُفترض حفظ هذا الملف في نفس مجلد ado-file (فمثلاً يتم حفظه في مجلد C:\ado\personal) الذي يحتوي على ملف *multicat.ado*.

أي ملف نصي يتم حفظه في أي مجلد معروف لبرنامج ستاتا بأنه مجلد خاص بملفات ado-file ويكون الملف النصي المحفوظ بتنسيق *filename.sthlp* سوف يتم عرضه على الشاشة بواسطة ستاتا عند طباعة الأمر **help** ***filename***. فمثلاً قد نقوم بكتابة الأمر أدناه في نافذة Do-file Editor وحفظ هذا الملف في المجلد C:\ado\personal باسم *multicat.sthlp*، ثم نقوم بطباعة الأمر **help multicat** في أي وقت، فإن برنامج ستاتا سوف يعرض النص.

```
multicat -- Multiple bar charts of categorical
variables}
multicat varlist [aw = weightvar] [if exp] [in
range],
      [missing] [by(groupvar)]

Description
multicat draws horizontal bar charts showing
percentages of
categorical variables. It saves one chart for
each of the variables in varlist. Graphs are
saved in the current default directory, with
file names based on variable names preceded by
a hyphen, such as -vote.gph or -region-
vote.gph. They are saved both in Stata's .gph
format and one other graphical file format
(.emf, .eps or .pdf) depending on operating
system.

Using analytical weights [aw = weightvar] with
multicat will result in percentages equivalent
to those obtained by svy: tab applied to data
declared as survey type, by a command such as
svyset [pw = weightvar]. The svy: prefix cannot
be used with multicat itself. Chapter 14 in
Statistics with Stata
(2012) has examples and discussion of multicat.

multicat requires that catplot is installed.
Type
```

```
findit catplot for instructions on installing
this unofficial
program, written by Nicholas Cox.

Options
missing includes missing values in the bar
chart and calculated percentages.

by(groupvar) draws an image containing separate
small charts for each value of groupvar.

Examples

multicat party wrongtrack vote

multicat party-vote [aw = weightvar], miss

multicat party-vote [aw = weightvar],
by(region)

References

Hamilton, Lawrence C. 2012.
Statistics with Stata. Belmont, CA: Ceng
```

ملفات المساعدة المفيدة هي تلك التي تحتوي على روابط ونصوص منسّقة ومربعات حوار وميزات أخرى يمكن تصميمها باستخدام Stata Markup and Control Language (SMCL). كل ملفات المساعدة الرسمية ببرنامج ستاتا، وملفات التسجيل، والنتائج المعروضة على الشاشة تقوم باستخدام SMCL، والتنسيق المرغوب، لملفات المساعدة بصفة عامة موجود في دليل المستخدم *User's Guide*.

النص، أدناه عبارة عن نسخة من SMCL لملف مساعدة خاص بـ**multicat** تتبع تقريباً توجيهات دليل المستخدم، وعند حفظ هذا الملف في مجلد ado\personal باسم *multicat.sthlp* ثم نقوم بطباعة الأمر **help multicat** سوف يظهر وكأنه ملف مساعدة رسمي.

```
{smcl}
{* *! version 2.0 21jul2012}{...}
{cmd:help multicat}
```

```
{hline}

{title:Title}
{phang}
{bf:multicat    --    Multiple    bar    charts    of
categorical variables}

{title:Syntax}
{p 8 17 12}
{cmd:multicat}    {it:varlist}    [{it:weight}]
[{cmd:if} {it:exp}]
[ { c m d : i n } { i t : r a n g e } ] { c m d
: , } [ { c m d a b : m i s s : i n g } ]
[{cmd:by(}{it:groupvar}{cmd:)}]

{title:Description}

{pstd}
{cmd:multicat}  draws  horizontal  bar  charts
showing percentages ofcategorical variables. It
saves one chart for each of the variablesin
{it:varlist}. Graphs are saved in the current
default  directory,with  file  names  based  on
variable names preceded by a hyphen, suchas
{it:-vote.gph} or {it:-region-vote.gph}. They
are saved both inStata's .gph format and one
other graphical file format (.emf, .epsor .pdf)
depending on operating system.

{pstd}
Using    analytical    weights    {cmd:[aw    =
}{it:weightvar}{cmd:]} with
{cmd:multicat}  will  result  in  percentages
equivalent to those
obtained by {cmd:svy: tab} applied to data
declared as survey type,by a command such as
{cmd:svyset [pw=}{it:weightvar}{cmd:]}.
The{cmd:svy:}  prefix  cannot  be  used  with
{cmd:multicat} itself. Chapter1 4 i n { b r o w
s e
"http://www.stata.com/bookstore/statistics-
with-stata/index.html":
Statistics with Stata} (2012) has examples and
discussion of {cmd:multicat}.
```

```
{pstd}{cmd:multicat} requires that
{cmd:catplot} is installed. Type{cmd:findit
catplot} for instructions on installing this
unofficialprogram, written by Nicholas Cox.

{title:Options}

{phang}
{cmdab:miss:ing} includes missing values in the
bar chart and
calculated percentages.

{phang}
{cmd:by(}{it:groupvar}{cmd:)} draws an image
containing separate
small charts for each value of {it:groupvar}.

{title:Examples}

{phang}
{cmd:. multicat party wrongtrack vote}

{phang}
{cmd:. multicat party-vote [aw = weightvar],
miss}

{phang}
{cmd:. multicat party-vote [aw = weightvar],
by(region)}

{title:References}

{pstd}
Hamilton, Lawrence C. 2012.
{browse
"http://www.stata.com/bookstore/statistics-
with-stata/index.html":
Statistics with Stata}. Belmont, CA:
Cengage.{p_end}
```

ملف المساعدة يبدأ بـ {smcl} الذي يأمر برنامج ستاتا باعتبار الملف من نوع SMCL، الأقواس المعكوفة { } تتضمن رموز SMCL والعديد منها له تنسيق {command:text} أو تنسيق {command arguments:text} والأمثلة أدناه توضح كيفية تفسير هذه الرموز.

{cmd:help multicat} يعرض النص "help multicat" كأمر، حيث يعرض "help multicat" بأي لون وحروف الخط يتم عرضها بشكل مناسب للأمر.

{hline} يرسم خطاً أفقياً.

{title:Title} يعرض النص "Title" كعنوان.

{phang} يقوم بإدراج مسافة بادئة للفقرة النصية التالية.

{bf:multicat- ...} يعرض النص بخط عريض.

{p 8 17 12} يقوم بتنسيق النص التالي كفقرة مع مسافة بادئة بمقدار 8 حروف والسطور التالية مع مسافة بادئة 17 حرفاً والهامش الأيمن يتم تضييقه بمقدار 12 حرفاً.

{it:varlist} يعرض النص *varlist* بخط مائل.

{cmdab:miss:ing} يعرض كلمة "missing" كأمر مع جعل حروف "miss" كأقل اختصار.

{browse "http://www.stata.com/bookstore/statistics-with-stata/index.html":Statistics...}

يربط النص "Statistics with Stata" مع موقع (URL) الإنترنت http://www.stata.com/bookstore/statistics-with-stata/index.html

حيث إن النقر على "Statistics with Stata" سوف يفتح متصفح الإنترنت على هذا الرابط.

دليل المستخدم *Programming Manual* يوفر تفاصيل أكثر عن استخدام هذه الأوامر، والعديد من أوامر SMCL الأخرى.

محاكاة مونت كارلو : Monte Carlo Simulation

محاكاة مونت كارلو تقوم بإنشاء وتحليل العديد من عينات البيانات الوهمية، مما يسمح للباحثين بالتحقق من سلوك تقنياتهم الإحصائية في المدى الطويل. الأمر **simulate** يجعل عملية تصميم المحاكاة واضحة وسهلة، حيث إنها تتطلب عددًا بسيطاً من البرامج الإضافية. في هذا الجزء، سوف يتم عرض مثالين عن ذلك.

عند البداية مع المحاكاة نحتاج إلى تعريف البرنامج الذي يقوم بإنشاء عينة واحدة من بيانات عشوائية ثم يحللها ويحفظ النتائج ذات العلاقة في الذاكرة. الملف التالي يُعرّف برنامج r-class (قادر على حفظ نتائج ( )r) باسم **meanmedian** هذا البرنامج يقوم عشوائياً بإنتاج 100 قيمة للمتغير *x* من توزيع طبيعي معياري، ثم يقوم بإنتاج 100 قيمة للمتغير *w* من توزيع طبيعي ملوّث: N(0,1) مع احتمالية 0.95 و N(0,10) مع احتمالية 0.05، التوزيعات الطبيعية الملوّثة يتم استخدامها في العادة في دراسات المتانة لمحاكاة المتغيرات التي تحتوي على أخطاء شاذة عرضية، بالنسبة لكلا المتغيرين، فإن **meanmedian** يقوم بحساب المتوسطات وقيم الوسيط.

```
* Creates a sample containing n=100
  observations of variables x
and w.
* x~N(0,1)                          x is standard
  normal
* w~N(0,1) with p=.95, w~N(0,10) with p=.05 w
  is contaminated
normal
* Calculates the mean and median of x and w.
* Stored results: r(xmean) r(xmedian) r(wmean)
  r(wmedian)
program meanmedian, rclass
    version 12.1
    drop _all
    set obs 100
    generate x = rnormal()
    summarize x, detail
    return scalar xmean = r(mean)
    return scalar xmedian = r(p50)
    generate w = rnormal()
```

```
   replace w = 10*w if runiform() < .05
   summarize w, detail
   return scalar wmean = r(mean)
   return scalar wmedian = r(p50)
end
```

ولأننا عرّفنا **meanmedian** كأمر r-class مثل **summarize** الذي يمكنه أن يحفظ نتائجه في قيم عددية ( )r، **meanmedian** يقوم بإنشاء أربع قيم عددية *r(xmean)* و *r(xmedian)* لمتوسط ووسيط المتغير، ونفس الشيء لمتوسط ووسيط المتغير *w* حيث تكون *r(wmean), r(wmedian)*.

عند تعريف **meanmedian** سواءً باستخدام do-file أو ado-file أو من خلال طباعة الأوامر بشكل تفاعلي، يمكننا أن نستخدم هذا البرنامج مرة أخرى من خلال الأمر **simulate**، ولإنشاء بيانات جديدة تحتوي على المتوسطات وقيم الوسيط للمتغير *x* والمتغير *y* من 5,000 عينة عشوائية نقوم بطباعة الأمر التالي:

```
.simulate xmean = r(xmean) xmedian = r(xmedian)
   wmean  =  r(wmean)  wmedian  =  r(wmedian),
reps(5000): meanmedian

    command: meanmedian
      xmean: r(xmean)
    xmedian: r(xmedian)
      wmean: r(wmean)
    wmedian: r(wmedian)

Simulations (5000)
```

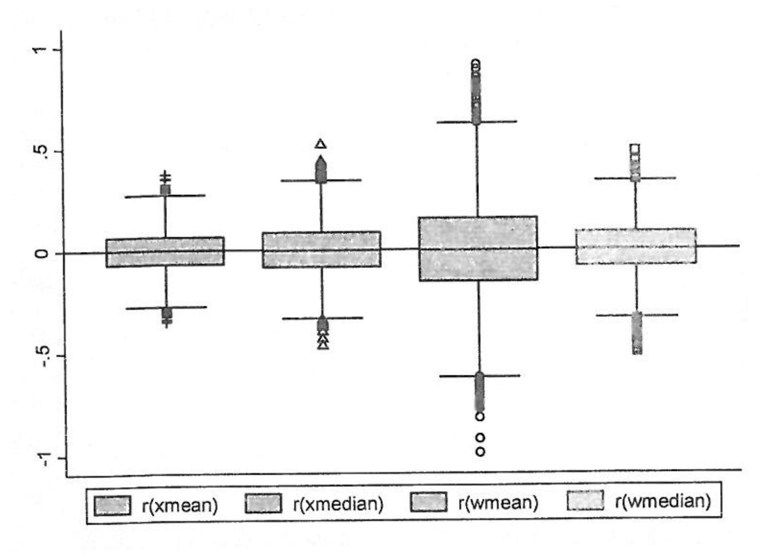
هذا الأمر يقوم بإنشاء المتغيرات *xmean, xmedian, wmean, wmedian* بناءً على نتائج ( )r من كل تكرار لـ **meanmedian**

```
.describe
Contains data
  obs:         5,000                        simulate: meanmedian
 vars:             4                        27 Mar 2014 22:50
 size:        80,000
------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------
xmean           float  %9.0g                  r(xmean)
xmedian         float  %9.0g                  r(xmedian)
wmean           float  %9.0g                  r(wmean)
wmedian         float  %9.0g                  r(wmedian)
------------------------------------------------------------------------
Sorted by:
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
xmean	5000	-.0023419	.1011693	-.3493929	.3718835
xmedian	5000	-.0014787	.1251765	-.4686761	.5143458
wmean	5000	-.002392	.2431499	-.994046	.905769
wmedian	5000	.0010861	.1282183	-.5034871	.4793077

المتوسطات لهذه المتوسطات وقيم الوسيط خلال 5000 عينة كلها تقريباً قريبة للصفر، وتتوافق مع توقعاتنا بأن متوسط ووسيط العينة يُفترض أن يوفرا تقديرات غير متحيّزة لمتوسطات المجتمع الصحيح (0) للمتغير x والمتغير w. وكما هو متوقع نظرياً، فإن المتوسط يبدو أقل تبايناً من الوسيط من عينة لأخرى عند تطبيقه على متغير x موزع توزيعاً طبيعياً، الانحراف المعياري للمتغير *xmedian* يساوي 0.125 وهو أكبر بدرجة كبيرة من الانحراف المعياري (0.101). ومن ناحية أخرى، فعند تطبيقه على متغير w غير طبيعي أو متغير يتأثر بالقيم المتطرفة، فإن العكس يكون صحيحاً: الانحراف المعياري للمتغير *wmedian* أقل بكثير من الانحراف المعياري للمتغير *wmean* (0.128 مقابل 0.244)، تجربة مونت كارلو توضح بأن الوسيط يبقى مقياساً مستقرًا نسبياً للمركز بالرغم من وجود القيم المتطرفة الشاذة في التوزيع الملوّث، بينما المتوسط ينخفض ويتنوع بشكل كبير من عينة لأخرى، الشكل (6.14) يعرض مقارنة بيانية للشكل الصندوقي (وبالصدفة يعرض كيفية التحكم في أشكال الصندوق البيانية وعلامات القيم المتطرفة)، ولإنشاء مسافة لأربعة متغيرات في مربع شرح لصف واحد، فسوف نقوم برسم الرموز بنصف حجمها الواقعي ((symxsize(*.5)).

```
.graph box xmean xmedian wmean wmedian,
 yline(0)
    legend(row(1) symxsize(*.5))
    marker(1, msymbol(+)) marker(2,
    msymbol(Th))
    marker(3, msymbol(Oh)) marker(4,
    msymbol(Sh))
```



الشكل (6.14)

مثالنا التالي يوسّع التحقق إلى طرق الثقة جامعاً معاً عدة موضـوعات من هذا الكتاب. برنامج **regsim** يقـوم بإنـشاء 100 مـشاهدة للمتغيـر x (الطبيعي القياسي) ومتغيرين اثنين لـ y حيث إن $y1$ هو دالة خطية للمتغير x زائداً الأخطاء الطبيعية المعيارية، $y2$ هو أيضاً دالة خطية للمتغير x ولكن يضيف الأخطاء الطبيعية الملوّثة، هذه المتغيرات تساعد في اكتشاف سـلوك عدة طرق انحدار في وجود توزيعات طبيعية وغير طبيعية لها أخطـاء ذات خطأ ذو انحراف كبير في منحنى التوزيع الطبيعي، تم استخدام أربع طـرق هي:

المربعات الصغرى العاديـة (**regress**)، والانحـدار الموثـوق (**rreg**)، والانحدار الربيعي (**qreg**)، والانحدار الربيعـي مـع الأخطـاء المعياريـة المُتكيّـسة bootstrapped standard errors (الأمـر **bsqreg** 500 تكـرار)؛ الموثوقية والانحدار الربيعي (الفصل 8). نظرياً يجب أن يُثبت مقاومة أكثر لتأثيرات القيم المتطرفة، ومن خلال تجربة محاكاة مونت كارلو نقوم باختبار ما إذا كان ذلك صحيحاً، الأمر **regsim** يقوم بتطبيق كل طريقـة لانحـدار

المتغير *y1* على المتغير *x* ثم بعد ذلك على انحدار المتغير *y2* على المتغيـــر *x*، وفي هذا التمرين سوف يتم تعريف البرنامج عن طريـــق ملـــف ado-file المُسمى *regsim.ado* والذي تم حفظه في المجلد ado\personal.

```
program regsim, rclass
* Performs one iteration of a Monte Carlo
  simulation comparing
* OLS regression (regress) with robust (rreg)
  and quantile
* (qreg and bsqreg) regression. Generates one n
  = 100 sample
* with x ~ N(0,1) and y variables defined by
  the models:
*
* MODEL 1: y1 = 2x + e1 e1 ~ N(0,1)
*
* MODEL 2: y2 = 2x + e2 e2 ~ N(0,1) with p =
  .95
* e2 ~ N(0,10) with p = .05
*
* Bootstrap standard errors for qreg involve
  500 repetitions.
*
version 12.1
if "`1'" == "?" {
global S_1 "b1 b1r se1r b1q se1q se1qb ///
b2 b2r se2r b2q se2q se2qb"
exit
}
drop _all
set obs 100
generate x = rnormal()
generate e = rnormal()
generate y1 = 2*x + e
reg y1 x
return scalar B1 = _b[x]
rreg y1 x, iterate(25)
return scalar B1R = _b[x]
return scalar SE1R = _se[x]
qreg y1 x
return scalar B1Q = _b[x]
return scalar SE1Q = _se[x]
bsqreg y1 x, reps(500)
return scalar SE1QB = _se[x]
replace e = 10 * e if runiform() < .05
```

```
generate y2 = 2*x + e
reg y2 x
return scalar B2 = _b[x]
rreg y2 x, iterate(25)
return scalar B2R = _b[x]
return scalar SE2R = _se[x]
qreg y2 x
return scalar B2Q = _b[x]
return scalar SE2Q = _se[x]
bsqreg y2 x, reps(500)
return scalar SE2QB = _se[x]
end
```

برنامج r-class يقوم بحفظ تقديرات الأخطاء المعيارية والمعاملات من ثمانية تحليلات انحدار، والنتائج يتم إعطاؤها أسماء مثل:

r(B1) معامل من انحدار OLS للمتغير *y1* على المتغير *x*.

r(B1R) معامل من الانحدار الموثوق للمتغير *y1* على المتغير *x*

r(SE1R) الخطأ المعياري لمعامل الثقة من النموذج 1

وهكذا، كل انحدارات الربيعات وانحدارات الثقة تتضمن تفاعلات متعددة: عادة من خمسة إلى عشرة تفاعلات للأمر **rreg** وحوالي خمسة للأمر **qreg** وعدة آلاف للأمر **bsqreg** مع 500 إعادة تقدير متكيّسة لحوالي خمسة تفاعلات لكل عينة، ولهذا فإن أمراً تنفيذياً واحدًا للبرنامج *regsim* يقوم بتحديد أكثر من 2,000 انحدار، والأمر أدناه يحدد خمسة تكرارات تتطلب أكثر من 10,000 انحدار.

```
.simulate b1 = r(B1) b1r = r(B1R) se1r = r(SE1R)
     b1q = r(B1Q) se1q = r(SE1Q) se1qb = r(SE1QB)
     b2 = r(B2)
     b2r = r(B2R) se2r = r(SE2R) b2q = r(B2Q)
     se2q = r(SE2Q)
     se2qb = r(SE2QB), reps(5): regsim
```

قد تريد إعادة تشغيل محاكاة بسيطة مثل التي قمنا بها أعلاه كتجربة لمعرفة الزمن المطلوب على جهاز الكمبيوتر لديك. وعموماً، فإننا قد نحتاج إلى زمن أطول للتجارب الأكبر؛ ملف البيانات *regsim.dta* يحتوي على نتائج من تجربة تضمنت 500 تكرار لبرنامج **regsim** وهي أكثر من مليون

انحدار، مُعاملات الانحدار وتقديرات الخطأ المعياري التي تم الحصول عليها من هذه التجربة تم تلخيصها في الجدول أدناه.

```
.describe

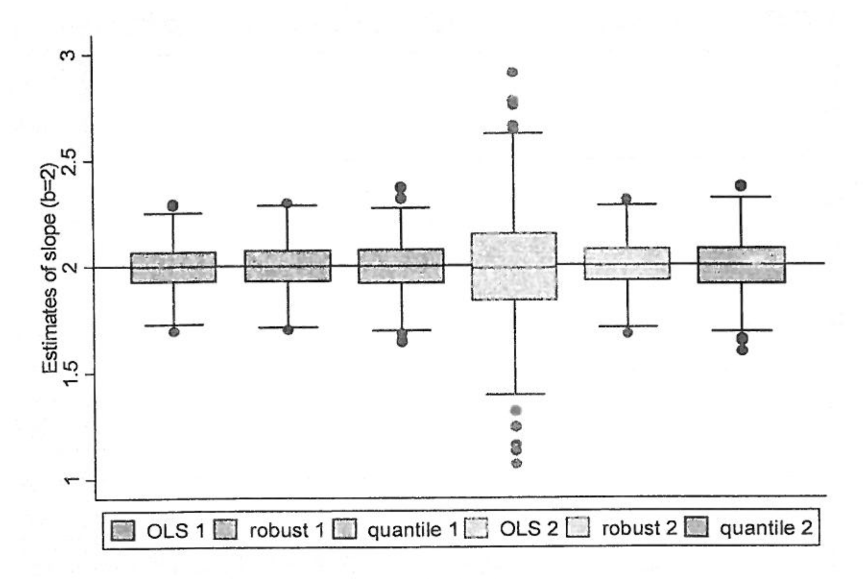
Contains data from C:\data\regsim.dta
  obs:          500                          Monte Carlo estimates of b in 500
                                               samples of n=100
 vars:           12                          2 Jul 2012 06:11
 size:       24,000
--------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
--------------------------------------------------------------------------------
b1              float  %9.0g                  r(B1)
b1r             float  %9.0g                  r(B1R)
se1r            float  %9.0g                  r(SE1R)
b1q             float  %9.0g                  r(B1Q)
se1q            float  %9.0g                  r(SE1Q)
se1qb           float  %9.0g                  r(SE1QB)
b2              float  %9.0g                  r(B2)
b2r             float  %9.0g                  r(B2R)
se2r            float  %9.0g                  r(SE2R)
b2q             float  %9.0g                  r(B2Q)
se2q            float  %9.0g                  r(SE2Q)
se2qb           float  %9.0g                  r(SE2QB)
--------------------------------------------------------------------------------
Sorted by:

.summarize

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          b1 |       500     1.99586     .104467   1.692893   2.293595
         b1r |       500    1.996901    .1077322   1.698501   2.294482
        se1r |       500    .1046559    .0108543   .0789753   .1523494
         b1q |       500     1.99658    .1246462   1.638891   2.370703
        se1q |       500      .13393    .0206363   .0801532   .2059937
-------------+--------------------------------------------------------
       se1qb |       500    .1373001    .0321417   .0532386   .2581546
          b2 |       500    1.986367    .2604184   1.066318    2.90484
         b2r |       500    1.997187    .1127494   1.674992   2.307488
        se2r |       500    .1087309    .0117741    .081064   .1564037
         b2q |       500    1.996925    .1314325   1.591606   2.370703
-------------+--------------------------------------------------------
        se2q |       500    .1416007    .0212944   .0880669   .2220859
       se2qb |       500    .1456451    .0343871   .0560117   .2704635
```

الشكل (7.14) يعرض توزيعات المُعامِلات كرسم صندوقي. ولجعل مربع شرح الرسم قابلاً للقراءة، سوف نقوم باستخدام الخيارات legend(symxsize(*.3) colgap(*.3)) والذي يقوم بتوسيع الرموز والفراغات بين الأعمدة في مربع شرح الرسم، بحيث تظهر بحجم 30% من حجمها الافتراضي، الأمر help legend option والأمر help relativesize يعرضان معلومات أكثر عن هذه الخيارات.

```
. graph box b1 b1r b1q b2 b2r b2q,
ytitle("Estimates of slope (b=2)")
yline(2)        legend(row(1)         symxsize(*.3)
colgap(*.3)
label(1  "OLS  1")  label(2  "robust  1")  label(3
"quantile 1")
label(4  "OLS  2")  label(5  "robust  2")  label(6
"quantile 2"))
```



الشكل (7.14)

نماذج الانحدار الثلاثة (OLS والموثوق والربيعي) أنتجت معامل متوسط يقوم بتقدير النموذجين، وهذه التقديرات ليست مختلفة معنوياً عن القيمة الصحيحة $\beta = 2$ ، ويمكن تأكيد هذا من خلال اختبارات t مثل:

```
.ttest b2r = 2
```

```
One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     b2r |     500    1.997187    .0050423    .1127494    1.987281    2.007094
------------------------------------------------------------------------------
    mean = mean(b2r)                                              t =  -0.5578
Ho: mean = 2                                     degrees of freedom =      499

    Ha: mean < 2                 Ha: mean != 2                 Ha: mean > 2
 Pr(T < t) = 0.2886         Pr(|T| > |t|) = 0.5772          Pr(T > t) = 0.7114
```

كل نماذج الانحدار تعطي تقديرات غير متحيزة لـ β، ولكن تبايناتها وكفاءتها تختلف من عينة لأخرى. وعند تطبيقها على نموذج الأخطاء الطبيعية 1 فإن OLS يُثبت بأنه الأكثر فعالية، وذلك كما توقعت نظرية جاوس ماركوف Gauss-Markov، كما أن الانحراف المعياري المشاهد لمُعاملات OLS يساوي 0.1045 مقارنة مع 0.0177 للانحدار الموثوق و0.1246 للانحدار الربيعي، والكفاءة النسبية والتي تعرض التباين المشاهد لمعامل OLS كنسبة من تباين مُقدّر آخر تعتبر طريقة معيارية لمقارنة مثل هذه الإحصائيات:

```
. quietly summarize b1
. scalar Varb1 = r(Var)
. quietly summarize b1r
. display 100*(Varb1/r(Var))
94.030265

. quietly summarize b1q
. display 100*(Varb1/r(Var))
70.242519
```

الحسابات أعلاه تستخدم نتيجة تباين r(Var) من الأمر **summarize**، سوف نقوم أولاً بالحصول على تباين تقديرات OLS للمتغير *b1* ثم نجعل تلك القيمة قيمة عددية *varb1*، بعد ذلك يتم الحصول على تباينات التقديرات الموثوقة *b1r* وتقديرات الربيعات *b1q* ومقارنتها مع *Varb1*. هذا يوضح أن الانحدار الموثوق كان تقريباً 94% كفء مثل كفاءة OLS عند تطبيقه على نموذج الأخطاء الطبيعية وهو قريب من نسبة كفاءة العينة الكبيرة 95% وهذه هي الطريقة التي يُفترض أن تكون موثوقة نظرياً (Hamilton 1992a)،

وفي المقابل، فإن الانحدار الربيعي يحقق كفاءة نسبية قدرها 70% تقريباً مع نموذج الأخطاء الطبيعية.

حسابات مماثلة لنموذج الأخطاء الملوثة يعطي صورة مختلفة، حيث إن OLS كان أفضل (الأكثر كفاءة) مُقدّر مع الأخطاء الطبيعية، ولكنه قد يكـون الأسوأ مع الأخطاء الملوثة.

```
. quietly summarize b2
. scalar Varb2 = r(Var)
. quietly summarize b2r
. display 100*(Varb2/r(Var))
533.47627

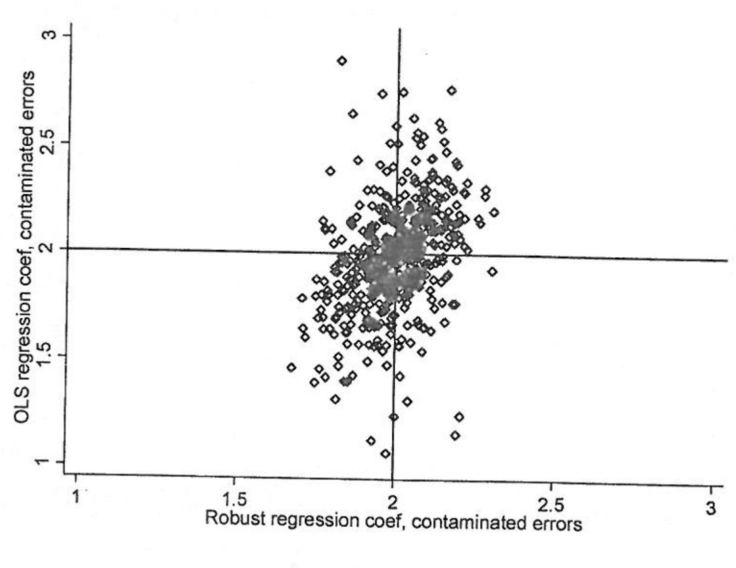
. quietly summarize b2q
. display 100*(Varb2/r(Var))
392.58875
```

القيم المتطرفة في نموذج الأخطاء الملوثة يؤدي بأن تكـون تقـديرات معاملات OLS متباينة بشكل كبير من عينة لأخرى. والذي يمكن ملاحظتـه بوضوح في الصندوق الرابع بالشكل (7.14)، تباين مُعـامِلات OLS أكبـر خمس مرات من التباين المتعلق بالمُعاملات الموثوقة، وأكبر بأربع مـرات تقريباً من المُعامِلات الربيعية. وبعبارة أخرى فقد تـم إثبـات أن الانحـدار الربيعي، والانحدار الموثوق كلاهما أكثر استقراراً من OLS في وجود القيم المتطرفة معطياً في المقابل أخطاء معيارية أقـل وفتـرات ثقـة أضـيق، والانحدار الموثوق يتفوّق على الانحدار الربيعي مع نماذج الأخطاء الطبيعية ونماذج الأخطاء الملوّثة.

الشكل (8.14) يعرض مقارنة بين OLS والانحدار الموثـوق عارضـاً شكل انتشار لــ 500 زوج من معاملات الانحدار. معاملات OLS (المحور العمودي) يتباين أكثر بكثير حول القيمة الصحيحة "2.0" أكثر من معـاملات rreg (المحور الأفقي).

```
. graph twoway scatter b2 b2r, msymbol(dh)
  ylabel(1(.5)3, grid)
      yline(2) xlabel(1(.5)3, grid gmin gmax)
      xline(2)
```

```
ytitle("OLS regression coef, contaminated
errors")
xtitle("Robust regression coef,
contaminated errors")
```



الشكل (8.14)

كما أن تجربة مونت كارلو تعرض معلومات حول الأخطاء المعيارية المقدرة في ظل كل طريقة وكل نموذج. متوسط الأخطاء المعيارية المقدرة يختلف عن الانحراف المعياري المشاهد للمُعاملات. والاختلاف بين الأخطاء المعيارية الموثوقة بسيط نسبياً، حيث إنه أقل من 4%، ونظرياً فإن اختلافات الأخطاء المعيارية الربيعية أكبر، حيث إنها 7% تقريباً، التقديرات المقبولة الصغرى تبدو أخطاء ربيعية متكيّسة تم الحصول عليها بواسطة الأمر **bsqreg**، ومتوسطات الأخطاء المعيارية المتكيّسة تفوق الانحراف المعياري المشاهد للمتغير *b1q* والمتغير *b2q* بحوالي 10 أو 11%، ويبدو أن التكيّس يُبالغ في تقدير التباين من عينة لأخرى.

محاكاة مونت كارلو أصبحت إحدى الطرق الرئيسة في البحوث الإحصائية الحديثة، كما أنها تلعب دوراً متزايدًا في التعليم الإحصائي، هذه الأمثلة توضح بعض الطرق السهلة لمعرفة طريقة استخدامها.

برمجة المصفوفات مع Mata : Matrix Programming with Mata

لغة برمجة المصفوفات ببرنامج ستاتا تسمى Mata، وتم شرحها بالتفصيل في إصدارين بدليل المستخدم *Mata Matrix Programming*، هذا الموضوع الضخم لا يقع ضمن نطاق هذا الكتاب، ولكن يتناسب مع هذا الفصل الختامي، حيث سنلقي نظرة سريعة على Mata، حيث إن أدواته البرمجية تفتح مجالات جديدة لتطوير ستاتا.

وبدلاً من قضاء فترة طويلة في شرح مفاهيم Mata ومميزاتها، سوف نبدأ مباشرة مع أمثلة عن كيفية كتابة برنامج يقوم بحساب انحدار المربعات الصغرى OLS، حيث إن نموذج الانحدار البسيط هو:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{u}$$

حيث إن **y** هي (n×1) موجّه عمودي لقيم المتغير التابع، **X** هي (n×k) مصفوف تحتوي على قيم (عادة) k-1 متغيرات تنبؤية وعمود 1، **u** هي (n×1) موجه الأخطاء، **b** هي (k×1) موجه مُعامِلات الانحدار ويتم تقديرها كما يلي:

$$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'y}$$

هذه الطريقة لحساب المصفوفة مألوفة لأجيال من طلبة الإحصاء، وتعتبر نقطة بداية مفيدة لمشاهدة كيفية عمل Mata.

ملف البيانات *reactor.dta*، يتضمن معلومات حول تكاليف إيقاف تشغيل خمسة مفاعلات طاقة نووية تم إقفالها في الفترة من 1968–1982، الميزة التعليمية في هذا المثال هي صغر حجم المصفوفات، حيث يمكن كتابتها بسهولة على السبورة أو الورق في حالة الحاجة إلى ذلك (e.g., Hamilton 1992a:340)، فسوف تتم معرفة كيف أن تكاليف إيقاف التشغيل قد تكون لها علاقة مع قدرة المفاعل وسنوات التشغيل.

```
. use C:\data\reactor.dta, clear
. describe
```

```
Contains data from C:\data\reactor.dta
  obs:            5                           Reactor decommissioning costs (from
                                                Brown et al. 1986)
 vars:            6                           2 Jul 2012 06:11
 size:          110
--------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
--------------------------------------------------------------------------------
site            str14  %14s                   Reactor site
decom           byte   %8.0g                  Decommissioning cost, millions
capacity        int    %8.0g                  Generating capacity, megawatts
years           byte   %9.0g                  Years in operation
start           int    %8.0g                  Year operations started
close           int    %8.0g                  Year operations closed
--------------------------------------------------------------------------------
Sorted by:  start
```

وبالطبع، فإن حساب انحدار OLS مع برنامج ستاتا سهل جـداً، حيـث وجدنا أن تكاليف إيقاف التشغيل لهذه الخمسة مفاعلات زادت بحوالي 0.176 مليون دولار (175,874 دولارًا) لكل ميجاوات قدرة على التوليد، وحـوالي 3.9 مليون دولار لكل سنة تشغيل. هذان المتغيـران التنبؤيـان يوضـحان حوالي 99% من تباين تكاليف إيقاف التشغيل (R^2_a=0.9895).

. *regress decom capacity years*

```
      Source |       SS       df       MS              Number of obs =       5
-------------+------------------------------           F(  2,      2) =  189.42
       Model |  4666.16571     2  2333.08286           Prob > F      =  0.0053
    Residual |  24.6342883     2  12.3171442           R-squared     =  0.9947
-------------+------------------------------           Adj R-squared =  0.9895
       Total |      4690.8     4      1172.7           Root MSE      =  3.5096

------------------------------------------------------------------------------
       decom |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    capacity |   .1758739   .0247774     7.10   0.019     .0692653    .2824825
       years |   3.899314   .2643087    14.75   0.005     2.762085    5.036543
       _cons |  -11.39963   4.330311    -2.63   0.119    -30.03146     7.23219
------------------------------------------------------------------------------
```

ملف ado-file أدناه يُعرّف برنامج ols0 مستخدماً أوامـر Mata، وهـو ببساطة يقوم بحساب موجه معاملات الانحدار b، في هذا المثال أوامر Mata تبدأ بـ :mata (هناك عدة طرق أخرى لاستخدام هذه الأوامر تفاعلياً أو في برامج تم شرحها في دليل المستخدم)، أول أمـرين اثـنين :mata يقومـان بتعريف الموجّه y والمصفوفة X كمعاينة للبيانات في الذاكرة وتم تحديـدها بواسطة كل المتغيرات في الطرف الأيسر (lhs) والمتغيرات فـي الطـرف الأيمن (rhs) التي تظهر سطر الأمر ols0، الثابت 1 يشكل آخر عمـود فـي المصفوفة X، كما أن ols0 يسمح بالمحـددات in أو if أو القـيم المفقـودة، والمعادلة المقدّرة هي:

$$b = (X'X)^{-1} X'y$$

ويتم كتابتها بلغة Mata على الشكل التالي:

```
mata: b = invsym(X'X)*X'y
```

أمر :mata الرابع يعرض محتويات النتائج للمتغير b

```
*! 21jun2012
*! L. Hamilton, Statistics with Stata (2012)
program ols0
   version 12.1
   syntax varlist(min=1 numeric) [in] [if]
   marksample touse
   gen cons_ = 1
   tokenize `varlist'
   local lhs "`1'"
   mac shift
   local rhs "`*'"
   mata: st_view(y=., ., "`lhs'", "`touse'")
   mata: st_view(X=., ., (tokens("`rhs'"),
   "cons_"), "`touse'")
   mata: b = invsym(X'X)*X'y
   mata: b
   drop cons_
end
```

وعند تطبيق ols0 على بيانات إيقاف تشغيل المفاعلات النووية فإن معاملات الانحدار التي تم الحصول عليها تشبه تلك التي تم الحصول عليها سابقاً بواسطة الأمر regress.

```
. ols0 decom capacity years
                          1

              1    .1758738974
              2    3.899313867
              3   -11.39963279
```

عند استخدام إصدارات Mata للمعادلات المعيارية، فإن برنامج ols1 (في الصفحة التالية) يقوم بإضافة حساب الأخطاء المعيارية وإحصائيات t واحتمالات اختبار t، ومرة أخرى نرى أن الحسابات تقود إلى نفس النتائج التي رأيناها سابقاً مع الأمر regress، الفواصل في آخر عبارة mata ببرنامج ols1 هي عبارة عن عوامل وتعني "قم بضم الأعمدة للمصفوفات التالية".

```
*! 21jun2012
*! L. Hamilton, Statistics with Stata (2012)
program ols1
   version 12.1
   syntax varlist(min=1 numeric) [in] [if]
   marksample touse
   gen cons_ = 1
   tokenize `varlist'
   local lhs "`1'"
   mac shift
   local rhs "`*'"
   mata: st_view(y=., ., "`lhs'", "`touse'")
   mata:   st_view(X=.,   .,   (tokens("`rhs'"),
   "cons_"), "`touse'")
   mata: b = invsym(X'X)*X'y
   mata: e = y - X*b
   mata: n = rows(X)
   mata: k = cols(X)
   mata: s2 = (e'e)/(n-k)
   mata: V = s2*invsym(X'X)
   mata: se = sqrt(diagonal(V))
   mata:    (b,    se,    b:/se,    2*ttail(n-k,
   abs(b:/se)))
   drop cons_
end
. ols1 decom capacity years
```

	1	2	3	4
1	.1758738974	.0247774037	7.098156835	.0192756353
2	3.899313867	.26430873	14.75287581	.0045631637
3	-11.39963279	4.330310729	-2.632520735	.1190686843

يمكننا أن نوسع هذا البرنامج، بحيث يحفظ النتائج ويعرضها في جداول ذات تنسيق أفضل يشبه تلك التي يعرضها الأمر **regress**، البرنامج **ols2** (في الصفحة التالية) يقوم بشيء مختلف حتى يوضح كيف أن Mata تضم المصفوفات معاً، حيث يقوم بدمج النتائج الرقمية المعروضة أعلاه في مصفوفات نصية تحتوي على عناوين للأعمدة، وقائمة بأسماء المتغيرات المستقلة، ويتم القيام بذلك من خلال أوامر **mata** إضافية، أحد هذه الأوامر يُعرّف موجه الصف **vnames_** والذي يحتوي على قائمة بأسماء المتغيرات. الفواصل في هذا الأمر تقوم بضم ثلاث مجموعات من الأعمدة: (1) العبارة "Yvar:" يتبعها اسم المتغيرات بالطرف الأيسر، (2) أسماء كل متغيرات الطرف الأيمن، (3) العبارة "_cons"

```
mata:vnames_="Yvar:`lhs'",tokens("`rhs'"),"_cons"
```

أمر **mata** الطويل التالي يقوم باستخدام محدد التعليقات الموجودة ضمن سطر الأمر وهي */ و/* حتى يستطيع Mata قراءة ما قبل نهاية آخر سطرين ويعتبر هذا كله كأمر واحد:

```
mata: vnames_', ("Coef." \ strofreal(b)),  /*
   */ ("Std. Err." \ strofreal(se)),       /*
   */  ("t"  \  strofreal(t)),  ("P>|t|"  \
     strofreal(Prt))
```

الأمر يعرض مصفوفة بها الصف الأول عبارة عن أسماء **vnames_** (وهو عمود أسماء المتغيرات). عمود أسماء المتغيرات مدمج باستخدام فاصلة مع موجه العمود الثاني الذي تم إنشاؤه مع كلمة "Coefs" كصف أول، أما بقية الصفوف فتم تعبئتها بمعاملات **b** والتي تم تحويلها من أرقام حقيقية إلى حروف، أما استخدام الشرطة الخلفية "\" فإنها تدمج الصفوف إلى مصفوفة مثلما تفعل الفاصلة "," التي تقوم بدمج الأعمدة. التحويل من أرقام

إلى حروف لقيم b مهم لجعل أنواع المصفوفات متوافقة، وهنـــاك عمليــات ممثلة في ols2 من الأعمدة الموصوفة للأخطـــاء المعياريـــة وإحـــصائيات t والاحتمالات.

```
*! 21jun2012
*! L. Hamilton, Statistics with Stata (2012)
program ols2
  version 12.1
  syntax varlist(min=1 numeric) [in] [if]
  marksample touse
  gen cons_ = 1
  tokenize `varlist'
  local lhs "`1'"
  mac shift
  local rhs "`*'"
  mata: st_view(y=., ., "`lhs'", "`touse'")
  mata:   st_view(X=.,   .,   (tokens("`rhs'"),
"cons_"), "`touse'")
  mata: b = invsym(X'X)*X'y
  mata: e = y - X*b
  mata: n = rows(X)
  mata: k = cols(X)
  mata: s2 = (e'e)/(n-k)
  mata: V = s2*invsym(X'X)
  mata: se = sqrt(diagonal(V))
  mata: t = b:/se
  mata: Prt = 2*ttail(n-k, abs(b:/se))
  mata:      vnames_      =      "Yvar:      `lhs'",
tokens("`rhs'"), "_cons"
  mata: vnames_', ("Coef." \ strofreal(b)), /*
  */ ("Std. Err." \ strofreal(se)), /*
  */   ("t"    \    strofreal(t)),    ("P>|t|"    \
strofreal(Prt))
  drop cons_
end
```

```
. ols2 decom capacity years
```

	1	2	3	4	5
1	Yvar: decom	Coef.	Std. Err.	t	P>|t|
2	capacity	.1758739	.0247774	7.098157	.0192756
3	years	3.899314	.2643087	14.75288	.0045632
4	_cons	-11.39963	4.330311	-2.632521	.1190687

تمارين Mata – مثلها مثل الأمثلة الأخرى في هذا الفصل – تعطي لمحة عن البرمجة في ستاتا. كما أن مجلة ستاتا *Stata Journal* تقوم بنشر تطبيقات أكثر توسعاً. وكل تحديث لبرنامج ستاتا يتضمن ملفات ado-files جديدة ومطوّرة، كما أن Online NetCourses تعرض طرقاً إرشادية لتعليمك كيف تكتب برامجك الخاصة بك.

مصادر البيانات

Dataset Sources

المنشورات أو صفحات الإنترنت أدناه، تعرض بعض المعلومات. وهذه المعلومات مثل: التعريفات، والمصادر الأصلية للبيانات، وعرض أوسع عن البيانات التي تم استخدامها في أمثلة هذا الكتاب. في الغالب، فـإن بيانـات الأمثلة هي مقتطفات من ملفات أكبر، أو تحتوي على متغيرات تم دمجها من أكثر من مصدر واحد. انظر **قائمة المراجع** للحصول على قائمة كاملة بها.

aids.raw

aids.dta

Selvin (1995)

Alaska_places.dta

Hamilton et al. (2011)

Alaska_regions.dta

Hamilton and Lammers (2011)

Antarctic2.dta

Milke and Heygster (2009)

Arctic9.dta

Sea ice extent: NSIDC (National Snow and Ice Data Center), Sea Ice Index.

http://nsidc.org/data/seaice_index/

Sea ice volume: PIOMAS (Pan-Arctic Ice Ocean Modeling and Assimilation System),

Polar Science Center, University of Washington. Arctic Sea Ice Volume Anomaly.

http://psc.apl.washington.edu/wordpress/research/projects/arctic-sea-ice-volume-ano

maly/

Annual air temperature anomaly 64–90 °N: GISTEMP (GISS Surface Temperature

Analysis), Goddard Institute for Space Studi e s , NASA.

http://data.giss.nasa.gov/gistemp/

attract2.dta

Hamilton (2003)

Canada1.dta

Canada2.dta

Federal, Provincial and Territorial Advisory Committee on Population Health (1996)

Climate.dta

NCDC global temperature: National Climatic Data Center, NOAA. Global Surface

Temperature Anomalies. http://www.ncdc.noaa.gov/cmb-faq/anomalies.php

NASA global temperature: GISTEMP (GISS Surface Temperature Analysis), Goddard

Institute for Space Studies, NASA. http://data.giss.nasa.gov/gistemp/

UAH global temperature: Universicy of Alabama, Huntsville.

http://vortex.nsstc.uah.edu/data/msu/t2lt/uahncdc.lt

Aerosol Optical Depth (AOD): Sato et al. (1993). Goddard Institute for Space Studies,

NASA. Forcings in GISS Climate Model. http://data.giss.nasa.gov/modelforce/strataer/

Total Solar Irradiance (TSI): Fröhlich (2006). Physikalisch-Meteorologischen

Observatoriums Davos, World Radiation Center (PMOD WRC). Solar Constant.

http://www.pmodwrc.ch/pmod.php?topic=tsi/composite/SolarConstant

Multivariate ENSO Index (MEI): Wolter and Timlin (1998). Earth Systems Research

Laboratory, Physical Sciences Division, NOAA. Multivariate ENSO Index.

http://www.esrl.noaa.gov/psd/enso/mei/mei.html

Global average marine surface CO2: Masarie and Tans (1995). Earth System Research

Laboratory, Global Monitoring Division, NOAA. Trends in Atmospheric Carbon

Dioxide. http://www.esrl.noaa.gov/gmd/ccgg/trends/global.html#global_data

election_2004i.dta

Robinson (2005). Geovisualization of the 2004 Presidential Election.

http://www.personal.psu.edu/users/a/c/acr181/election.html

electricity.dta

California Energy Commission (2012). U.S. Per Capita Electricity Use by State, 2010.

http://energyalmanac.ca.gov/electricity/us_per_capita_electricity-2010.html

global1.dta

global2.dta

global3.dta

global_yearly.dta

Multivariate ENSO Index (MEI): see *climate.dta*

NCDC global temperature: see *climate.dta*

Granite2011_6.dta

Hamilton (2012). Also see "Do you believe the climate is changing?" by Hamilton (2011)

athttp://www.carseyinstitute.unh.edu/publications/IB-Hamilton-Climate-Change-National-

NH.pdf

Greenland_sulfate.dta

Mayewski, Holdsworth et al. (1993); Mayewski, Meeker et al. (1993). Also see Sulfate and

Nitrate Concentrations at GISP2 from 1750–1990.

http://www.gisp2.sr.unh.edu/DATA/SO4NO3.html

Greenland_temperature.dta

GISP2 ice core temperature: Alley (2004). NOAA Paleoclimatology Program and World

Data Center for Paleoclimatology, Boulder.

ftp://ftp.ncdc.noaa.gov/pub/data/paleo/icecore/greenland/summit/gisp2/isotopes/gisp2_temp

_accum_alley2000.txt

Summit temperature 1987–1999: Shuman et al. (2001)

greenpop1.dta

Hamilton and Rasumssen (2010)

GSS_2010_SwS.dta

Davis et al. (2005). National Opinion Research Center (NORC), University of Chicago.

General Social Survey. http://www3.norc.org/GSS+Website/

heart.dta

Selvin (1995)

lakewin1.dta

lakewin2.dta

lakewin3.dta

lakesun.dta

lakesunwin.dta

Lake Winnipesaukee ice out: http://www.winnipesaukee.com/index.php?pageid=iceout

Lake Sunapee ice out: http://www.town.sunapee.nh.us/Pages/SunapeeNH_Clerk/

Also see Hamilton et al. (2010a) at

http://www.carseyinstitute.unh.edu/publications/IB_Hamilton_Climate_Survey_NH.pdf

MEI0.dta

MEI1.dta

Multivariate ENSO Index: see *climate.dta*

MILwater.dta

Hamilton (1985)

Nations2.dta

Nations3.dta

Human Development Reports, United Nations Development Program. International Human

Development Indicators. http://hdrstats.undp.org/en/tables/

oakridge.dta

Selvin (1995)

planets.dta

Beatty (1981)

PNWsurvey2_11.dta

Hamilton et al. (2010b, 2012). Also see "Ocean views" by Safford and Hamilton (2010),

at http://www.carseyinstitute.unh.edu/publications/PB_Safford_DowneastMaine.pdf

reactor.dta

Brown et al. (1986)

shuttle.dta

shuttle0.dta

Report of the Presidential Commission on the Space Shuttle Challenger Accident (1986)

Tufte (1997)

smoking1.dta

smoking1.dta

Rosner (1995)

snowfall.xls

Hamilton et al. (2003)

southmig1.dta

southmig2.dta

Voss et al. (2005)

student2.dta

Ward and Ault (1990)

whitemt1.dta

whitemt2.dta

Hamilton et al. (2007)

writing.dta

Nash and Schwartz (1987)

قائمة المراجع

References

Albright, J.J. and D.M. Marinova. 2010. "Estimating Multilevel Models Using SPSS, Stata, and

SAS." Indiana University.

Alley, R.B. 2004. GISP2 Ice Core Temperature and Accumulation Data. IGBP PAGES/World

Data Center for Paleoclimatology Data Contribution Series #2004-013. NOAA/NGDC

Paleoclimatology Program, Boulder CO, USA.

Beatty, J.K., B. O'Leary and A. Chaikin (eds.). 1981. *The New Solar System*. Cambridge, MA:Sky.

Belsley, D.A., E. Kuh and R.E. Welsch. 1980. *Regression Diagnostics: Identifying InfluentialData and Sources of Collinearity*. New York: John Wiley & Sons.

Box, G.E.P., G.M. Jenkins and G.C. Reinsel. 1994. *Time Series Analysis: Forecasting andControl*. 3rd ed. Englewood Cliffs, NJ: Prentice–Hall.

Brown, L.R., W.U. Chandler, C. Flavin, C. Pollock, S. Postel, L. Starke and E.C. Wolf. 1986.*State of the World 1986*. New York: W. W. Norton.

California Energy Commission. 2012. "U.S. per capita electricity use by state in 2010."

http://energyalmanac.ca.gov/electricity/us_per_capita_electricity-2010.html
accessed 3/13/2012

Chambers, J.M., W.S. Cleveland, B. Kleiner and P.A. Tukey. 1983. *Graphical Methods forData Analysis*. Belmont, CA: Wadsworth.

Chatfield, C. 2004. *The Analysis of Time Series: An Introduction*, 6th edition. Boca Raton, FL:CRC.

Cleveland, W.S. 1993. *Visualizing Data*. Summit, NJ: Hobart Press.

Cleves, M., W. Gould, R. Gutierrez and Y. Marchenko. 2010. *An Introduction to SurvivalAnalysis Using Stata*, 3rd edition. College Station, TX: Stata Press.

Cook, R.D. and S. Weisberg. 1982. *Residuals and Influence in Regression*. New York:Chapman & Hall.

Cook, R.D. and S. Weisberg. 1994. *An Introduction to Regression Graphics*. New York: JohnWiley & Sons.

Cox, N.J. 2004a. "Stata tip 6: Inserting awkward characters in the plot." Stata Journal4(1):95–96.

Cox, N.J. 2004b. "Speaking Stata: Graphing categorical and compositional data." *Stata Journal*4(2):190–215.

Davis, J.A. T.W. Smith and P.V. Marsden. 2005. General Social Surveys, 1972–2004Cumulative File [computer data file]. Chicago: National Opinion Research Center [producer].Ann Arbor, MI: Inter-University Consortium for Political and Social Research [distributor].

Diggle, P.J. 1990. *Time Series: A Biostatistical Introduction*. Oxford: Oxford University Press.

Enders, W. 2004. *Applied Econometric Time Series*, 2nd edition. New York: John Wiley &Sons.

Everitt, B.S., S. Landau and M. Leese. 2001. *Cluster Analysis*, 4th edition. London: Arnold.

Federal, Provincial and Territorial Advisory Commission on Population Health. 1996. *Reporton the Health of Canadians*. Ottawa: Health Canada Communications.

Foster, G. and S. Rahmstorf. 2011. "Global temperature evolution 1979–2010." *EnvironmentalResearch Letters* 6. DOI:10.1088/1748-9326/6/4/044022

Fox, J. 1991. *Regression Diagnostics*. Newbury Park, CA: Sage Publications.

Fröhlich, C. 2006. "Solar irradiance variability since 1978—Revision of the PMOD compositeduring solar cycle 21." *Space Science Review* 125:53–65.

Gould, W., J. Pitblado and B. Poi. 2010. Maximum Likelihood Estimation with Stata, 4th edition. College Station, TX: Stata Press.

Hamilton, D.C. 2003. "The Effects of Alcohol on Perceived Attractiveness." Senior Thesis.Claremont, CA: Claremont McKenna College.

Hamilton, J.D. 1994. *Time Series Analysis. Princeton*, NJ: Princeton University Press.

Hamilton, L.C. 1985. "Who cares about water pollution? Opinions in a small-town crisis."*Sociological Inquiry* 55(2):170–181.

Hamilton, L.C. 1992a. *Regression with Graphics: A Second Course in Applied Statistics*. PacificGrove, CA: Brooks/Cole.

Hamilton, L.C. 1992b. "Quartiles, outliers and normality: Some Monte Carlo results." Pp. 92–95in J. Hilbe (ed.) *Stata Technical Bulletin Reprints*, Volume 1. College Station, TX: Stata Press.

Hamilton, L.C., D.E. Rohall, B.C. Brown, G. Hayward and B.D. Keim. 2003. "*Warming wintersand New Hampshire's lost ski areas: An integrated case study.*" *International Journal ofSociology and Social Policy* 23(10):52–73.

Hamilton, L.C., B.C. Brown and B.D. Keim. 2007. "Ski areas, weather and climate: Time seriesmodels for New England case studies." *International Journal of Climatology* 27:2113–2124.

Hamilton, L.C. and R.O. Rasmussen. 2010. "Population, sex ratios and development inGreenland." *Arctic* 63(1):43–52.

Hamilton, L.C., B.D. Keim and C.P. Wake. 2010a. "Is New Hampshire's climate warming?"New England Policy Brief No. 4. Durham, NH: Carsey Institute, University of New Hampshire.

Hamilton, L.C., C.R. Colocousis and C.M. Duncan. 2010b. "Place effects on environmentalviews." *Rural Sociology* 75(2):326–347.

Hamilton, L.C. and R.B. Lammers. 2011. "Linking pan-Arctic human and physical data." *PolarGeography* 34(1–2):107–123.

Hamilton, L.C., D.M. White, R.B. Lammers and G. Myerchin. 2011. "Population, climate andelectricity use in the Arctic: Integrated analysis of Alaska community data." *Population andEnvironment* 33(4):269–283. DOI: 10.1007/s11111-011-0145-1.

Hamilton, L.C. 2012. "Did the Arctic ice recover? Demographics of true and false climate facts." Paper presented at the American Sociological Association. Denver, Colorado, August17–20.

Hamilton, L.C., T.G. Safford and J.D. Ulrich. 2012. "In the wake of the spill: Environmentalviews along the Gulf Coast. *Social Science Quarterly* DOI: 10.1111/j.1540-6237.2012.00840.x

Hardin, J. and J. Hilbe. 2012. *Generalized Linear Models and Extensions*, 3rd edition. CollegeStation, TX: Stata Press.

Hoaglin, D.C., F. Mosteller and J.W. Tukey (eds.). 1983. *Understanding Robust andExploratory Data Analysis*. New York: John Wiley & Sons.

Hoaglin, D.C., F. Mosteller and J.W. Tukey (eds.). 1985. *Exploring Data Tables, Trends andShapes*. New York: John Wiley & Sons.

Hosmer, D,W., Jr., S. Lemeshow and S. May. 2008. *Applied Survival Analysis: RegressionModeling of Time to Event Data*, 2nd edition. New York: John Wiley & Sons.

Hosmer, D.W., Jr. and S. Lemeshow. 2000. *Applied Logistic Regression*, 2nd edition. NewYork: John Wiley & Sons.

Kline, R.B. 2010. *Principles and Practice of Structural Equation Modeling*, Third Edition. NewYork: Guilford.

Korn, E.L. and B.I. Graubard. 1999. *Analysis of Health Surveys*. New York: Wiley.

Lean, J.L. and D.H. Rind. 2008. "How natural and anthropogenic influences alter global andregional surface temperatures: 1889 to 2006." Geophysical Research Letters 35DOI:10.1029/2008GL034864

Lee, E.T. 1992. *Statistical Methods for Survival Data Analysis*, 2nd edition. New York: JohnWiley & Sons.

Lee, E.S. and R.N. Forthofer. 2006. *Analyzing Complex Survey Data*, second edition. ThousandOaks, CA: Sage.

Levy, P.S. and S. Lemeshow. 1999. *Sampling of Populations: Methods and Applications*, 3rd Edition. New York: Wiley.

Li, G. 1985. "Robust regression." Pp. 281–343 in D. C. Hoaglin, F. Mosteller and J. W. Tukey(eds.) *Exploring Data Tables, Trends and Shapes*. New York: John Wiley & Sons.

Long, J.S. 1997. *Regression Models for Categorical and Limited Dependent Variables*.Thousand Oaks, CA: Sage Publications.

Long, J. S. and J. Freese. 2006. *Regression Models for Categorical Dependent Variables UsingStata*, 2nd edition. College Station, TX: Stata Press.

Luke, D.A. 2004. *Multilevel Modeling*. Thousand Oaks, CA: Sage.

Mallows, C.L. 1986. "Augmented partial residuals." *Technometrics* 28:313–319.

Masarie, K.A. and P.P. Tans. 1995. "Extension and integration of atmospheric carbon dioxidedata into a globally consistent measurement record." *Journal of Geophysical Research*100:11593–11610.

Mayewski, P.A., G. Holdsworth, M.J. Spencer, S. Whitlow, M. Twickler, M.C. Morrison, K.K.Ferland and L.D. Meeker. 1993. "Ice-core sulfate from three northern hemisphere sites: Sourceand temperature forcing implications." *Atmospheric Environment* 27A(17/18):2915–2919.

Mayewski, P.A., L.D. Meeker, S. Whitlow, M.S. Twickler, M.C. Morrison, P. Bloomfield, G.C.Bond, R.B. Alley, A.J. Gow, P.M. Grootes, D.A. Meese, M. Ram, K.C. Taylor and W. Wumkes.1994. "Changes in atmospheric circulation and ocean ice cover over the North Atlantic duringthe last 41,000 years." *Science* 263:1747–1751.

McCullagh, P. and J.A. Nelder. 1989. *Generalized Linear Models*, 2nd edition. London:Chapman & Hall.

McCulloch, C.E. and S.R. Searle. 2001. *Generalized, Linear, and Mixed Models*. New York:Wiley.

Milke, A., and G. Heygster. 2009. "Trend der Meereisausdehnung von 1972–2009." TechnicalReport, Institute of Environmental Physics, University of Bremen, August 2009, 41 pages.

http://www.iup.uni-bremen.de/iuppage/psa/documents/Technischer_Bericht_Milke_2009.pdf

Mitchell, M.N. 2008. *A Visual Guide to Stata Graphics*, 2nd edition. College Station, TX: StataPress.

Mitchell, M.N. 2012. *Interpreting and Visualizing Regression Models Using Stata*. CollegeStation, TX: Stata Press.

Moore, D. 2008. *The Opinion Makers: An Insider Reveals the Truth about Opinion Polls*.Boston: Beacon Press.

Nash, J. and L. Schwartz. 1987. "Computers and the writing process." *CollegiateMicrocomputer* 5(1):45–48.

Rabe–Hesketh, S. and B. Everitt. 2000. *A Handbook of Statistical Analysis Using Stata*, 2nd edition. Boca Raton, FL: Chapman & Hall.

Rabe-Hesketh, S. and A. Skrondal. 2012. *Multilevel and Longitudinal Modeling Using Stata*,3rd edition. College Station, TX: Stata Press.

Raudenbush, S.W. and A.S. Bryk. 2002. *Hierarchical Linear Models: Applications and DataAnalysis Methods*, 2nd edition. Newbury Park, CA: Sage.

Raudenbush, S.W., A.S. Bryk, Y.F. Cheong & R. Congdon. 2005. HLM 5: *HierarchicalLinear and Nonlinear Modeling*. Lincolnwood, IL: Scientific Software International.

Report of the Presidential Commission on the Space Shuttle Challenger Accident. 1986.Washington, DC.

Robinson, A. 2005. "Geovisualization of the 2004 presidential election." Available at http://www.personal.psu.edu/users/a/c/acr181/election.html (accessed 3/8/2008).

Rosner, B. 1995. *Fundamentals of Biostatistics*, 4th edition. Belmont, CA: Duxbury Press.

Safford, T.G. and L.C. Hamilton. 2010. "Ocean views: Coastal environmental problems as seenby Downeast Maine residents." New England Policy Brief No. 3. Durham, NH: Carsey Institute,University of New Hampshire.

Sato, M., J.E. Hansen, M.P. McCormick and J.B. Pollak. 1993. "Stratospheric aerosol opticaldepths, 1850–1990." Journal of Geophysical Research 98:22,987–22,994.

Selvin, S. 1995. *Practical Biostatistical Methods*. Belmont, CA: Duxbury Press.

Selvin, S. 1996. *Statistical Analysis of Epidemiologic Data*, 2nd edition. New York: OxfordUniversity.

Shuman, C.A., K. Steffen, J.E. Box and C.R. Stearns. 2001. "A Dozen years of temperatureobservations at the summit: Central Greenland automatic weather stations 1987–99." *Journalof Applied Meteorology*, 40:741–752.

Shumway, R.H. 1988. *Applied Statistical Time Series Analysis*. Upper Saddle River, NJ:Prentice–Hall.

Skrondal, A. and S. Rabe-Hesketh. 2004. *Generalized Latent Variable Modeling: Multilevel,Longitudinal, and Structural Equation Models*. Boca Raton, FL: Chapman & Hall/CRC.

StataCorp. 2011. *Getting Started with Stata for Macintosh*. College Station, TX: Stata Press.

StataCorp. 2011. *Getting Started with Stata for Unix*. College Station, TX: Stata Press.

StataCorp. 2011. *Getting Started with Stata for Windows*. College Station, TX: Stata Press.

StataCorp. 2011. *Mata Reference Manual* (2 volumes). College Station, TX: Stata Press.

StataCorp. 2011. *Stata Base Reference Manual* (3 volumes). College Station, TX: Stata Press.

StataCorp. 2011. *Stata Data Management Reference Manual*. College Station, TX: Stata Press.

StataCorp. 2011. *Stata Graphics Reference Manual*. College Station, TX: Stata Press.

StataCorp. 2011. *Stata Programming Reference Manual*. College Station, TX: Stata Press.

StataCorp. 2011. *Stata Longitudinal/Panel Data Reference Manual*. College Station, TX: StataPress.

StataCorp. 2011. *Stata Multivariate Statistics Reference Manual.* College Station, TX: StataPress.

StataCorp. 2011. *Stata Quick Reference and Index.* College Station, TX: Stata Press.

StataCorp. 2011. *Stata Structural Equation Reference Manual.* College Station, TX: Stata Press.

StataCorp. 2011. *Stata Survey Data Reference Manual.* College Station, TX: Stata Press.

StataCorp. 2011. *Stata Survival Analysis and Epidemiological Tables Reference Manual.*College Station, TX: Stata Press.

StataCorp. 2011. *Stata Time-Series Reference Manual.* College Station, TX: Stata Press.

StataCorp. 2011. *Stata User's Guide.* College Station, TX: Stata Press.

Street, J.O., R.J. Carroll and D. Ruppert. 1988. "A note on computing robust regressionestimates via iteratively reweighted least squares." *The American Statistician* 42(2):152–154.

Tufte, E.R. 1990. *Envisioning Information.* Cheshire CT: Graphics Press.

Tufte, E.R. 1997. *Visual Explanations: Images and Quantities, Evidence and Narrative.*Cheshire CT: Graphics Press.

Tufte, E.R. 2001. *The Visual Display of Quantitative Information*, 2nd edition. Cheshire CT:Graphics Press.

Tufte, E.R. 2006. *Beautiful Evidence.* Cheshire CT: Graphics Press.

Tukey, J.W. 1977. *Exploratory Data Analysis.* Reading, MA: Addison–Wesley.

Velleman, P.F. 1982. "Applied Nonlinear Smoothing," pp.141–177 in Samuel Leinhardt (ed.)*Sociological Methodology 1982.* San Francisco: Jossey-Bass.

Velleman, P.F. and D.C. Hoaglin. 1981. *Applications, Basics, and Computing of ExploratoryData Analysis.* Boston: Wadsworth.

Verbeke, G. and G. Molenberghs. 2000. *Linear Mixed Models for Longitudinal Data.* NewYork: Springer.

Voss, P.R., S. McNiven, R.B. Hammer, K.M. Johnson and G.V. Fuguitt. 2005. "County-specificnet migration by five-year age groups, Hispanic origin, race, and sex, 1990–2000." Ann Arbor,MI: Inter-university Consortium for Political and Social Research, 2005-05-23.

Ward, S. and S. Ault. 1990. "AIDS knowledge, fear, and safe sex practices on campus."*Sociology and Social Research* 74(3):158–161.

White, J.W.C., R.B. Alley, J. Brigham-Grette, J.J. Fitzpatrick, A.E. Jennings, S.J. Johnsen, G.H.

Miller, R.S. Nerem and L. Polyak. 2010. "Past rates of climate change in the Arctic."Quaternary Science Reviews 29(15–16):1716–1727.

Wild, M., A. Ohmura and K. Makowski. 2007. "Impact of global dimming and brightening onglobal warming," *Geophysical Research Letters.* DOI:10.1029/2006GL028031.

Wolter, K. and M.S. Timlin. 1998. "Measuring the strength of ENSO events—how does1997/98 rank?" *Weather* 53:315–324.